DOCUMENT RESUME

ED 156 218 IR 006 451

AUTHOR Schermer, Diana
TITLE A Survey and Analysis of Educational Review Activities (A Preliminary Review). Product 3.
INSTITUTION Educational Products Information Exchange Inst., New York, N.Y.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 75
CONTRACT 400-75-0023
NOTE 93p.; For related documents, see IR 006 450-452

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS Decision Making; *Decision Making Skills; Educational Development; Educational Needs; Educational Resources; Educational Technology; *Evaluation Criteria; *Evaluation Methods; Guidelines; *Instructional Materials; Media Selection; *Selection

ABSTRACT

Despite the fact that reviews of educational products for selection purposes usually have to be based on evidence less conclusive than formal evaluations, these reviews must be approached as thoughtfully and as rationally as possible. More fruitful results will come from review procedures in which the purpose of and criteria for the review are clearly analyzed and stated. This document discusses the educational product review process in terms of: (1) types of reviews, (2) special issues about criteria, (3) problems with evidence, (4) types of selection criteria including need criteria, intrinsic quality, practicality and effectiveness, and (5) practicality of reviews. This is the third publication of an EPIE (Educational Products Information Exchange) project investigating the criteria used to select instructional materials. (Author/STS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED156218

NIE Products on Instructional Materials Selection
Created under National Institute of Education
(D.H.E.W.) Contract #400-75-0023
By EPIE Institute, 475 Riverside Drive, New York, N.Y. 10027

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Product 3. A SURVEY AND ANALYSIS OF EDUCATIONAL REVIEW ACTIVITIES
(A PRELIMINARY REVIEW)

- by Diana Schermer

prepared under the direction of Susan S. Klein

for: School Practice and Service Program
Office of Dissemination Resources
National Institute of Education
May 1975

IR006451

I. Introduction

An educational review, as the term is used in this paper, is an activity carried out by an organization that must choose one or several educational products from many possibilities. A school choosing films or textbooks, a board of education choosing materials for an approved list, an organization choosing materials for inclusion in a catalog of recommended materials, an agency choosing among newly developed products to decide where to use dissemination funds -- all of these are review activities. In a review, a choice must be made. Often the available data are inadequate, the judgments to be made have insufficient foundation, and the evidence concerning the effectiveness of the products is unconvincing; but still a choice must be made. Teaching cannot stop until sufficient evidence to base selections on is obtained.

Typically in a review, very limited time is spent in rating each product, usually a matter of hours. This factor distinguishes a review from an evaluation. For example, in an empirical evaluation, a staff designs an experiment, administers tests before and after a product is used, and collects and analyzes data to see if a product is effective. In short, an evaluation collects data about a product or products, while a review uses whatever data and information are available in order to make a decision among products.

In a few cases, it may be difficult to distinguish between a review and an evaluation. As one example, the review staff may be making the measurements and collecting the data needed to make a decision; thus an evaluation is done as part of the review. Due to the huge expense involved, this is fairly rare. Another example is a review performed to choose material to recommend and report on, by a consumer agency perhaps, when the decision process is less at issue than sum-

marizing the information on the quality of the product. This activity is indistinguishable from a summative evaluation in which only scanty evidence is gathered.

Given the nature of reviews, the rationale for this paper is as follows: In a review a decision must be made, even if the evidence is unconvincing, a needs assessment is lacking, and many of the judgments are based on value judgments or conjecture. We cannot simply throw up our hands saying that if we cannot go about this scientifically it does not matter how we do it. It does matter. For we cannot avoid choosing educational materials; we can only avoid choosing them irrationally. We must approach the review process as thoughtfully and as rationally as possible under these limitations. Value judgments, assumptions, and guesses must be as explicit and as well thought out as possible. Never should they enter into the process covertly through carelessly worded criteria. No matter how limited the information our decisions must be based on, more fruitful results will come from review procedures in which the purpose of, and criteria for, review are clearly thought through and stated.

II. Types of Reviews

Any discussion of specific issues raised in educational reviews must include the wide range of review activities that are performed with their differing purposes, conditions, emphases, and criteria. The range will be discussed in this section.

A. Single or mixed topic. Some reviews are used to choose among products with similar subject matter. A school, for instance, may need to select a film

on reproduction in plants, or a new biology textbook. Other reviews may be used in choosing among products with very different subjects, for preparation of a catalog, or for allocation of dissemination funds. The 1974 NIE Dissemination Review (14, p.17) covers products as diverse as teacher training material and a product designed for teaching English to Spanish-speaking children. Some single topic reviews could be quite broad, as in the case of a biology department's review for purposes of allocating funds among various materials, from films to kits to textbooks to complete "systems."

A single topic review can often compare products on one scale or set of scales if all the products are designed to do the same thing, and the main question of the review is how well they do it. Even if the purposes of the various products being reviewed are not identical, all of them may be used for the same general purpose, such as "improving our tenth-grade biology program"; this gives some basis for comparison. A mixed topic review raises a question which is as crucial as it is difficult: What priority do we give to the goals, effects, and needs addressed by each product? Mixed topic reviews must deal with this question of priority, for they cannot otherwise distinguish between high quality products meeting a crucial need and those relevant to an insignificant one.

B. Types of decisions. In general, reviews are used to make six kinds of decisions:

(1) To develop a product. A review for purposes of choosing between various proposals to allocate development funds raises issues that are fundamentally different from other educational reviews. First, there are neither a product nor evaluation data available for review, only the proposal. Second, the need for

the product becomes the main criterion. Third, the credentials of the developer become much more significant. Because of these three differences, reviews of this type are not included in this report.

(2) To continue development. In a review for the purpose of allocating funds for continuing the development of products, a product in some degree of completion is available, often with preliminary evaluation data. The unique imperative of these reviews is to distinguish between faults that can be corrected and those that cannot. Unclear language, unfair treatment of races or of the sexes, inadequate treatment of certain content areas, and evidence of effectiveness which is insufficient or based on faulty experimental design are examples of faults that may be corrected. The review should not eliminate products at this stage because of these inadequacies. Also, this kind of review has a secondary purpose in addition to making the decision, i.e., to provide the developer with suggestions relative to how and where the product may be improved -- a formative function analogous to that of a formative evaluation.

(3) To disseminate. A common type of review is one done by an agency to allocate funds for disseminating newly developed products. Because only limited funds are available, these reviews generally rank products in order of quality, rather than just decide between acceptable and unacceptable ones. A fine distinction must be made if, say, only enough funds are available for seven out of ten high-quality products. Dissemination reviews sometimes rate the dissemination plan for each product; this can be done as a separate stage of the review, after the products have been reviewed for quality. In other cases, the dissemination plans are negotiated by the producer and funding agency after the

review is completed.

National dissemination reviews are often completely mixed topic -- with no limitations on the topic whatsover -- and are thus faced with the problem of rating the priorities of products with very different objectives. Also, the products are likely to be used in many different schools in different sections of the country, raising questions of suitability in these varied situations and of what the national priorities are. This national perspective is quite different from that of one school choosing materials for its own students. General purposes must be considered rather than the specific needs of specific students. Finally, issues of marketability are important in dissemination reviews in two very divergent ways: An agency may want to allocate dissemination funds for products that are easily marketable by commercial publishers, with a view of getting maximum impact for their money. Or, they may want to support high priority products which commercial publishers would not be interested in because of a limited number of potential buyers, for example.

A commercial publisher deciding which products to publish is in essence carrying out a dissemination review. But because of the many special considerations emphasized by publishers, these reviews are not covered in this report. Some of those considerations are how to complete a line of products, how to open new markets, how to complement other products, or how to take advantage of a present market.

(4) To package a program. These reviews are designed to choose successful ongoing programs, such as a compensatory reading program, to be packaged as products. (See our definitions of program and product in the glossary, Appendix A.) In addition to reviewing the effectiveness of the programs, these reviews involve

a number of special considerations: whether the program could be easily packaged, whether the school is willing to cooperate with the project, whether the success of the program is dependent upon factors that are not included in packaging, e.g., charismatic leadership. Also, these reviews must consider that, since they are dealing with a potential product rather than one already finished, faults and shortcomings of the program as it is currently functioning may be corrected later. Thus, a fault that would eliminate a product might just be a required revision when found in a program to be packaged. In a sense, a decision to package a program is a special case of a dissemination decision. Due to the nature of these considerations, this kind of review turns out to be quite different in practice, and is, consequently, listed separately.

Since the programs under review are actually in use in a school, the evidence for their effectiveness is likely to be less convincing than that for a product that was field-tested for the purpose of collecting evaluation data. The controls, for example, are often statistical rather than experimental, making it difficult to ensure that the aspects of the program that would be packaged are the ones responsible for the effect. Often in these reviews the quality of the evidence becomes the main basis of discrimination between programs which are accepted and those which are rejected. In a report of a review of compensatory reading programs performed by the American Institute of Research, it was found that the "quality of local evaluation became essentially synonymous with the judged quality of the program" (2, p.4). The designers of a program review must consider whether they want to choose programs on this basis -- and must be especially sure that the design, wording, and weightings of criteria in their review do not favor programs shown conclusively to be minimally effective over programs shown less conclusively to be very effective -- simply because of the quality of their evaluation design or evidence.

(5) To recommend. Many reviews have the purpose of selecting materials to be recommended for use, as in a catalog of recommended American history products, for instance. State school boards preparing a list of approved materials are essentially performing this kind of review. However, they often look at only a few specified criteria, such as legal requirements -- whether a product has been learner verified and revised, for example -- rather than judge the overall quality of the products reviewed.

In practice, reviews to select recommended materials are often similar to dissemination reviews. Both types frequently emphasize effectiveness and overall quality and must somehow take into account the wide range of possible uses and circumstances for each product. Three main differences are: First, while a catalog can include as many products as are found worthy, dissemination funds may limit the number of products to fewer than the total number of worthwhile ones, requiring a much finer level of discrimination. A review for a catalog, for example, might screen out all the unacceptable products and include all the remaining ones in the catalog. Such an approach would usually not be possible in a dissemination review. Second, the catalog review need not face the problem of judging needs and priorities. The catalog can simply include all high quality materials regardless of the priority of the needs addressed by the product. Setting of priorities is, in turn, left to the users who choose materials for their own needs with the assistance of the catalog. Third, many catalog reviews simply screen out products on the basis of a few criteria and then include all the others.

Some of the catalog reviews have a dual purpose: first to select the best or appropriately screened products, second to summarize the results of their

findings for each product in the catalog. The summarizing function can be stated in terms of the overall quality of the products, in terms of verification of claims made by developers and of the evidence supporting the claims, or in terms of the content of the product. In some cases catalogs are not selective of material on the basis of the quality, and, hence, all relevant products are included. Such a catalog does not really involve a review process since emphasis is not placed on choice. Most of this paper will therefore not be relevant to these activities. An exception is the catalog that, although not selective, includes a summary of the quality of each product. Although a catalog of this type does not emphasize quality, it does require investigating the same criteria and issues that would make a choice possible.

Other catalogs involve varying amounts of review activities. Verifying the claims made for a product, or the evidence supporting a product, does not in itself constitute a review, although it is an important part of a review. Even the verification of the data indicating that the product attains its objectives does not constitute a review, although this too is certainly a major part of a review activity. A review must consider if the objectives are appropriate and worthwhile, consider any side effects unrelated to the objectives, consider the quality of production, the accuracy of the contents, and so on. A product whose objective, for example, is to raise students' score on a test, must still be rated in these other aspects before a review can be used to make a decision.

A catalog which simply summarizes the contents and characteristics of various products with no ratings of quality or verification of evidence has little in common with a review activity. This is not to say that such a catalog is useless. On the contrary, for many actual users this kind of summary is probably

the most useful information they can receive since teachers must find material which is appropriate to their needs before even considering the relative effectiveness of various products. Catalogs which verify claims and evidence as well as those which verify effectiveness are of great use also. Many users prefer making their own judgments of quality based on this information as opposed to simply accepting the judgments of the group producing a catalog or to receiving descriptions which exclude any reviewing information. These are considerations that producers of catalogs must address relative to how they can be most helpful in their particular situation. The point here is that review activities as analyzed in this report -- the process of making decisions or recommendations about products based on their judged quality -- have varying degrees of similarity to the production of catalogs.

(6) To use a product. Probably the most common kind of review is the one performed by a teacher, school, or school board in choosing materials which will be actually used. Although it may be done carefully and thoughtfully, or haphazardly, each time a study is done or a search is made to select the best or most appropriate material for a situation, it is, in essence, a review.

Several features distinguish these reviews from national-level dissemination or catalog reviews. For one thing, the product is being chosen for a specific purpose and specific situation. The criterion which we call appropriateness becomes crucial, that is, the question of whether what the product can do "fits" the needs of the school. For making this judgment, a thorough instructional materials-analysis instrument, such as one developed by EPIE Institute is helpful. Stated briefly, this instrument, EPIEform A, is designed to consider elements of instructional design -- Intents (developer's rationale, goals, and objectives); Contents (scope and sequence); Methodology (including the type of

learning environment, nature of learner interaction, teacher/learner roles, teaching/learning approaches, and organization of space, time, grouping, and pacing); and Means of Evaluation -- and, in addition, integrity of design, or, how these elements work, or do not work, together. Moreover, scrutiny is applied to physical description (technical and aesthetic quality), administrative/ teacher preparation requirements, community acceptance, fairness (in terms of sex roles, racism, and cultural, regional, and national backgrounds), accuracy, and currency. Third, user reviews must not only consider the strengths and weaknesses of different materials, but also the relative strengths of different combinations of materials. A comprehensive textbook must be rated against a more specific textbook supplemented with other materials, for example. Finally, the priority criterion, when it must be rated, comes up in a different form. A school board, for example, may have fixed allocated funds for social studies material, and may have to choose priorities between various texts, films, tapes, etc., that could be used with different schools or classes.

User reviews often have two stages: first to survey what is available and decide what is worthy of further investigation, and then to make the final decision.

C. Kind of material. Reviews differ in the kind of material that is being reviewed. We found three categories of material: products, programs, and practices. All the reviews, however, were of instructional material, that is, material which is designed to have an educational purpose. Not all materials used in education are instructional materials. A daffodil may be used in a biology class to great advantage, but that does not make it an instructional material, and we would not want to review its effectiveness, quality of production, or the relevance of its purposes. Such is also the case of a novel studied in an English class or a painting in an art class. On the other hand,

the lesson designed by the biology teacher and using the flower are instructional (being considered a "practice," as defined below), and could be evaluated or reviewed. Some cases are ambiguous: A short story is not instructional, but a collection of short stories might be; a play usually is not instructional, but a student's edition of it might be. Most of the criteria to be discussed in this paper will simply not apply to material that is not instructional.

Products include such materials as textbooks, workbooks, films, and tapes, as well as any set combination of these that constitute a course or what is called a program, such as a reading program or a mathematics program (a more flexible use of the word "program" than the way it is used in this report). A product is a transportable educational material. It can be held in your hands, bought and sold, and used in many different situations. Staff is never part of a product, though members of a staff may of course use a product.

A program is an instructional system ongoing in a specific location. Typically, a program will have products that it uses, a staff, a budget, standard procedures, general goals and specific objectives, and so on, that is, it has the characteristics of a system. Several of the reviews covered by this report are reviews of programs undertaken in order to find successful ones to be packaged and disseminated to other schools. These packages fall under our category of products rather than programs, since they are transportable. What is commonly called a reading program or a mathematics program by publishers is definitely included in our category of a product. We apologize for this confusion but it seems unavoidable if a clear differentiation between program and product is to be established.

The term "practice" has been used most indiscriminately. It commonly refers to any procedure that is used in a specific location. This includes pedagogical techniques, lessons, courses or units within courses, programs, as well as administrative practices. Except for the last one, all of them can be reviewed in a search for successful practices to be disseminated since they are all instructional. We would propose that the term "practice" be used in a more limited sense, to mean an educational procedure which is used at a specific location and which is not comprehensive or complete enough to be considered a program. A practice would thus be a standard procedure or a type of behavior. If it were made into a product to be disseminated, that product might well take the form of a teacher's handbook or manual.

Reviews of practices and programs have some major differences from those of products, since the former are ongoing in specific locations while the latter can be field-tested in several locations. These differences are discussed in the previous section, part B.4.

The major focus of this paper will be on product reviews. Although we will usually be speaking of products, most of the issues discussed refer equally to reviews of programs and practices.

D. Research vs. educational functions. Some reviews are not concerned solely with choosing what seems to be the best of alternative products. Rather, they are undertaken to identify products which have the best evidence of being good products, or the best evidence for being effective. Given the state of the art of educational research in general, and of educational evaluation in particular, and given the difficulty and expense of conducting a thorough evaluation, we cannot be at all sure that the best products will be the ones with the most

convincing evidence for effectiveness. The products with the best evidence, for example, will tend to be the ones whose goals are the most straightforward and easy to measure, and the ones which are used in situations where controls are easy to institute.

An organization conducting a review must consider carefully how much emphasis it wishes to put on the quality of evidence for a product. For it may intend to emphasize evidence of effectiveness, not solely as an aid for finding the best products, but also as a way to promote future research and evaluation. That is, it might want to choose products with good evidence over products which might actually be better ones but which lack credible evidence simply in order to pressure producers to put more effort toward gathering evidence in the future. This is the same process by which learner-based revision (formerly called "Learner Verification and Revision") is being encouraged: If products that have gone through L-BR are routinely chosen over ones that have not, soon all products would go through L-BR.

In studying the reports of various review activities, we found major differences between those which serve this function of promoting research and those which are simply looking for the best product to use in a particular situation or for a particular purpose. The former stress effectiveness criteria and the quality of evidence, often far more than could be justified by the usefulness of these criteria in finding the best products. They also stress uniqueness and originality, hoping to assure that new methods and principles are tested. On the other hand, reviews, usually by actual users, whose only concern is to choose the best product they can find -- they might be said to have solely an educational function -- are much more concerned with criteria of appropriateness, such as

whether the goals, contents, and educational strategy of the product fit in with the situation in which the product is to be used.

Any agency which wishes its reviews to serve this research function as well as the educational one must be cautious of a possible built-in bias from the over-emphasis of effectiveness and quality of evidence. This bias would be in favor of products with objectives that are straightforward and easy to measure, or even mechanical or trivial. For these are the products whose effectiveness is easiest to demonstrate; and they will have a built-in advantage in any review with a strong emphasis on evidence. An English composition book emphasizing spelling and grammar would have an advantage over the one emphasizing creative writing. We may prefer the first, but that should be a decision based on the judgment of the reviewers, and not subtly built into the review procedure itself.

The division between those reviews with a research function and those with an educational function alone is part of a wider split between educators and educational researchers. As far as review activities go, this split will only be narrowed if users are willing to pay more attention to the effectiveness and evidence, which they often dismiss as being of limited relevance, while research-oriented groups pay more attention to factors of appropriateness, which they often dismiss as too judgmental to be dealt with scientifically.

III. Special Issues about Criteria

In general, the criteria used in reviews can be divided into five categories: (1) the effectiveness of the product; (2) the need for the product; (3) the intrinsic quality, such as the accuracy of the contents; (4) the practicality, such as the cost; and (5) the quality of the supporting evidence. This section

will cover certain general issues about these categories. The range of specific criteria that should be considered in each category will be discussed in section V.

A. Screening criteria. Any criterion can be used in two ways, either as a screen for rejecting a product, or as a scale for making quantitative rating. For example, effectiveness can be a screening criterion if products failing to meet the criterion are automatically rejected. Or it could be used as a scaled criterion if each product in a review is given a rating for effectiveness, and this rating is combined with ratings on other criteria to reach a final decision.

The two uses of a criterion can be combined by using a cutoff point on a scale. Products could be rated on an effectiveness scale and the resultant rating could be combined with others so that any product falling below a certain rating could be automatically eliminated.

Very often a review will have several stages, including a screening stage that uses screening criteria for finding the relevant products, and scaled criteria to choose among them. Any time a criterion can be used as a screen, the subsequent review is greatly simplified. This is because this use avoids the need for giving a specific rating on that criterion, the need for justifying that specific rating, and the need to combine that rating, with appropriate weightings, with others.

Often, the whole question of need can be handled as a screening criterion. This is usually not possible in reviews of products that address different needs; in most of these cases the relative importance of the need must be rated. But need

criteria can be used as screens when the purpose of a review is to decide among products that all address the same need, for example, if a school board is searching for a biology text for the tenth grade. In such a case, products can be screened according to whether they address that need, and the subsequent review does not have to involve a quantitative rating of need.

One aspect of need criteria that is often crucial in reviews is the question of appropriateness or suitability. That is, part of the issue of whether the product's goals or effects are needed is the issue of how well these goals or effects fit in with ongoing programs, other products that are being used, and so on. For a teacher, school, or school board choosing products to use in their classes, this is often the key question. In many reviews the appropriateness criterion can be used as a screening criterion so that all products which are not appropriate can be eliminated and the review need only choose among the remaining ones. It would be a serious mistake, however, to use this as a screen and not a scaled criterion in deciding which product is the most appropriate, for the result could easily be the choice of a product that is very effective in doing something but not in doing exactly what is needed in that situation.

Effectiveness can sometimes be used as a screening criterion. In this case all products not meeting some standard of effectiveness would be eliminated in an initial stage of a review, and the choice would only be among a select group of products. Similarly, quality of evidence can occasionally be used as a screen. This has often been done in reviews designed for purposes of selecting programs to be packaged and disseminated, but has the disadvantage of eliminating such a high percentage of the programs or products that one wonders whether many programs that are actually most effective are being eliminated because of the difficulty in obtaining convincing evidence.

Various value questions can be covered as screening criteria. Some reviews may intend to screen out educational strategies, while others may wish to eliminate possible products on the basis of social fairness or other values. Finally, whether a product has been "L-BRed" is often a screening criterion in states where only products put through the L-BR process can be considered.

In all these cases, using criteria as screens can greatly simplify the subsequent stages of a review. But whenever criteria are used as screens and not also as scaled criteria, the final outcome of the review will ignore the question of how well those criteria are met, and the designers of the review must be aware that that question will be ignored.

B. The need-effectiveness question. At the most fundamental level, a review has just one question to answer: How well the product does do what the reviewers need. To complete the review, only a few supplementary criteria need be added: How much does it cost? Are there any reasons the product must be elimated, such as unfair treatment of women or minorities? Are the contents accurate? How reliable is the evidence? Other considerations are refinements of these questions.

The first refinement is to rate the need. We now have two questions: How important is the need? How well does the product meet the need? The former is the need question or priority question, the latter the effectiveness question. When criteria are categorized as need criteria or effectiveness criteria, that division is based on this level of refinement. Need criteria reflect the importance of the need, the goals, or the effects. The effectiveness criteria rate how well the needs or goals are met, or how strong the effects are.

A second refinement is possible once a product's goals are considered. The

need-effectiveness question can be broken into three parts: Does the product attain its goals? Are the goals adequate to meet the need? How important is that need? If the criteria of the review are worded in terms of goals, all three of these questions must be asked.

A final refinement is possible if we add consideration of objectives. Objectives, compared to goals, are more specific and are usually defined operationally. Now the need-effectiveness question can conceivably have four parts: Does the product attain its objectives? Are the objectives adequate for attaining the goals? Do the goals meet the need? How important is the need? This level of refinement is too detailed to be used in practice. For a product whose objectives seem inadequate, for example, it would be difficult for a reviewer to decide whether they were inadequate to meet the goals or whether the goals were inadequate to meet the need. This discrimination could only be made for a product which clearly states its objectives, goals, and the need it addresses. This final refinement is mentioned only to help clarify the need-effectiveness question in theory.

In general, a division into two questions, one for need and one for effectiveness, will be adequate for most reviews. The intermediate questions are not essential. But when the producers have given information about objectives, or if there are evaluation data on how well the product attains its objectives, it may be useful to use the second level of refinement.

The division of criteria into need criteria and effectiveness criteria is, in some ways, a fact-value distinction. First the reviewers decide what they need -- a value question -- and then they decide how well the product meets that need -- a factual question. This has encouraged some reviewers to emphasize effectiveness

criteria and consider need criteria to be unscientific. The problem here is that a rating of how well goals are attained or how well needs are met is useless without a rating of the importance of those needs or goals. This tendency has encouraged producers to simply state the goals that they know will be achieved by the product, rather than those things that need to be achieved. The result of this tendency can only be to further convince users that educational research and evaluation are of limited value to them.

C. Effectiveness. Effectiveness is a word that can have many meanings. To some extent, which definition is used in a review is less important than ensuring that the same definition is used throughout, and that it is clear to the reviewers which definition is being used. Otherwise, one reviewer, for example, may give a rating of how well a product achieves its goals, another of how powerful an effect the product is. Thus, the reliability of the review is destroyed. Since effectiveness and need are tied together in many ways, a clear definition of effectiveness is required so that an appropriate definition of need can be used in the review. If effectiveness, for example, is defined as how well a product achieves its goals, the most appropriate definition of need is how important or appropriate the goals are.

We have come across five possible definitions of effectiveness, some of which are fairly similar; however, they might give very different ratings for the same product. Each of these is meant to be a scaled criterion, and each could be broken into many, more detailed criteria, which will be discussed in section V.

(1) The goal definition: How well the product achieves its goals. This definition is especially appropriate in a review of products that have the same goals. Otherwise, it is essential that the relevance or importance of the goals also be rated as a need criterion. This is to avoid obtaining similar

overall ratings for two products, one of which achieves trivial goals, the other crucial goals. Even if the need for the goals is rated, the context for the rating has still been set by the producer's goals, which may or may not be the best context. For example, if the goals of the product are not really what is needed in the learning situation where it is likely to be used, the reviewer would want an effectiveness rating for how well the product does what is needed; a rating for how well it attains its goals is not very relevant.

(2) The effects definition: How strong an effect the product has. This definition makes obvious some of the dimensions of effectiveness that must be singled out: How strong is the effect on each student? How many students are affected? How broad or narrow is the range of effects on each student? These questions will be discussed in Section V.

As with the first definition, the need for these effects must be rated whenever the products covered in a review have different kinds of effects. Reviewers must take into account effects that are powerful but trivial. This definition raises a question of standards that must be noticed. If an effect is rated as "very strong," for example, it means either in comparison with other products, or, as compared to what is needed. In the first case an additional criterion of comparative effectiveness would be redundant; in the second, a vague assessment of need is being used without being made explicit, allowing the possibility that different reviewers will have different implicit ideas of what is needed.

(3) The operation definition: How well the product works using a specified operational standard of effectiveness. This approach is especially useful when the appropriate standard is evident. If the standard is controversial, however, the designers of the review will have a major effect on the outcome of the review. The product that does best by this standard has an advantage, even if

the reviewer thinks another product is more effective on a more appropriate standard.

(4) The need definition: How well the product fulfills the need. This definition suggests that each reviewer first try to formulate what is needed in the product, make those needs as explicit as possible, and rate the product on how effective it is in meeting the needs. If the review includes products addressing different needs, the importance of the needs would have to be rated as a need criterion. This is the approach taken by the model rating form proposed in Section VII. There reviewers are asked to state the objectives they think the product should have if different from the view of the producer. Ratings of effectiveness are then done in terms of those objectives.

This approach has the advantage of dealing directly with the underlying issue of effectiveness: Does the product do what is needed? The reviewer is required to make a judgment about what is needed and this nebulous area is brought right into the heart of the review with little likelihood of hard evidence or scientific methodology to help, given the present state of the art of educational needs assessment. Yet in a review that avoids this vague area in favor of areas where more scientific rigor is available there is the risk that a product with proved effectiveness will be used in a situation where some other effects are called for.

This approach, and also the following one, could be combined with any one of the first three. For example, if the first definition of effectiveness (the goal definition) is used, but it is found that the objectives are not very relevant to the need, asking how well the product does what <u>is</u> needed could be very helpful.

(5) The use definition: What the product can be used for and how effective

it is. This approach, which is in essence merely a refinement of the previous one, suggests the following procedure: Each reviewer first tries to see how the product relates to the needs of the situation, and then tries to decide what the product could best be used for. Effectiveness is then rated as how well the product does what the reviewer would expect of it under certain circumstances, and need is rated on how important that use is. The proposed use would be made explicit by each reviewer, and in a panel review, would be subject to discussion by the other reviewers.

This approach makes explicit the fact that a needs assessment as done in the fourth approach would not be very helpful in a review unless an assessment is done of the possible uses of the product. This is especially obvious when considering a product with a very limited goal. A thorough needs assessment, then, is less important than an examination of the possible uses of the product.

This approach addresses most directly the concerns of any review which has the purpose of finding the best product as opposed, say, to promoting evaluation or development methodologies. An educational product is a tool, and the crucial questions are: What can it do? How well does it do it?

In many reviews, the results obtained by some or all of these five definitions will be the same. For example, if the appropriate goals of a product are very obvious, then the goals of the developer and the needs of the situation as perceived by the reviewer might well be identical. The effects of the product might be to meet those needs. In this case, a reviewer would give the same rating regardless of definition used. But even when the end result of a review would be the same for all the definitions, reliability requires that the review be clear about which definition is being used, and that the need criteria are consistent with it.

Part of the immense complexity of review activities comes from the cases in which the goals of a product do not match the needs of the situation, or situations in which the effects investigated by the evaluators are not what is called for. Yet a decision must be made in these situations; teaching cannot stop until the proper evidence is available. In these situations it is most clear that the outcome of the review can be no more dependable than the judgment of the reviewer. The best one can hope for is that the reviewer is given a chance to use his or her judgment as clearly and rationally as possible. Even if the goals of the product are appropriate, and the evaluation evidence is directly relevant, whose judgment is to be used to decide that they are? If we want the review to express the reviewer's best judgment, he or she is the one who must make this key decision. And of the five definitions of effectiveness, the last two -- the needs definition and the use definition -- give the reviewer a chance to formulate the overall needs of the situation.

D. Need. Rating the priority of the need addressed by a product is a very difficult task, and many reviews are fortunately able to avoid it. In these reviews, all the products have the same goal, or they all address the needs of the situation, or the need for each of them is equal. The products could be screened in an initial state of the review to see that one of these conditions is met, and no further rating of need would be required. Catalog reviews, which intend to leave the choice of priorities entirely to the users, may also omit need criteria and choose products on the basis of other criteria. The catalog would then include both high-priority and low-priority products.

In many cases the need must be rated, however, no matter how difficult it is, and the rating must be given a high weighting in the final decision. One case is the review with no specific topic; when products address very different needs,

some might address crucial needs while others insignificant needs, and it may be important to choose only the former, especially if the allocation of funds depends on the results of the review. Another case is when priorities must be set as part of the review. The social studies department, for example, may have a certain amount of money allocated for new products, and must choose between products with different uses: some narrow, some broad, some designed for one target population. Setting the priorities must therefore be part of the review. A third case is the review designed to find materials for use in a certain class, school, or school system. Here the appropriateness of the products to ongoing programs using other products is often the most important issue of all. We consider appropriateness a need criterion because it deals with the question of how important the goals or effects of the product are.

The reviewers here are clearly being asked to make a value judgment. They must decide which of the desires of the consumers are legitimate needs and which are subjective preferences; they must decide which areas and skills should be emphasized and which should be neglected. A review that eliminates these issues misses the most important issues of all, and favor will be shown toward those products whose goals are most easily measured but may be the least significant.

Few reviews are likely to exclude the need issues in situations where they are crucial. What is common is to smuggle those issues into a section of the review where their presence is less conspicuous and embarrassing. A review may, for example, ask the reviewer to rate how powerful the effects of a product are, and mean by that, how <u>powerful</u> <u>and</u> <u>significant</u> the effects are. Or the reviewer may try to judge the educational significance of the effects as an effectiveness criterion.

We would recommend an opposite strategy. In many reviews, value judgments about needs are essential. These judgments should be as explicit and overt as possible, with explanations of the rationale behind them as part of the review. In a panel review, each reviewer could rate needs and give his or her explanations. These explanations could be examined as part of the panel discussion. In a review leading to a recommendation to be put in a catalog, the outcome of this discussion could be included in the catalog as essential information about the assumptions under which the review was carried out. This is the strategy used in the model rating form proposed in Section VII of this report.

The need criterion has at least as many possible versions as effectiveness: How high a priority is the need? Does the product actually meet the need? Are the objectives of the product what are actually needed? As with effectiveness, the choice of which version will be included in the review, although important, is often less crucial than being clear about which version is being used. Otherwise, reviewers may be giving ratings based on quite different conceptions of what the question is. The choice, however, should be made in consideration of which approach to effectiveness is being used.

In discussing the aspects of need appropriate for each version of effectiveness, one must keep in mind the overall need-effectiveness issue: how useful a product is. The answer will depend on three elements: (1) How strong an effect does it have? (2) How well does that effect meet the students' need? (3) How high is the priority of that need? Phrased in terms of objectives, the three questions become: (1) How well does the product meet its objectives? (2) How well do the objectives address the need? (3) How important is that need? The second and third questions, in either phrasing, are very different. A product could satisfy

a low-priority need very well; or it could address a high priority need but not meet that need very successfully.

In both phrasings, the first question is clearly one of effectiveness, the third of need; the second could be considered part of either category. If it is thought of in terms of the importance of the effects (or goals), it seems like a need criterion. As worded, on the other hand, it seems more like an effectiveness criterion. It probably matters little, provided all three questions are considered in the review, and it is clear to the reviewer which question is being asked.

One other possible confusion. We use the term "need criterion" to mean how important the need or problem addressed by the product is. This must not be confused with what the need for the product is or why the product is needed; these considerations will be descriptions, not ratings.

Each of the approaches to effectiveness mentioned earlier presents special issues to be considered under need:

(1) The first approach, how well the product achieves its goals or objectives, requires two questions be asked: How well do the goals or objectives address the need? How important is the need? Without these intermediate questions the review would not result in a distinction between two products; one of which attains important goals well, and another which attains trivial goals well. Often reviews have combined the last two questions into one: How important are the goals? Such a combination would mask the difference between a product with a specific purpose -- and which has goals that are too narrow to be rated important -- and one which has unimportant goals because they are not what is called for.

(2) The second approach to effectiveness, how strong an effect the product has, raises the same considerations as the first. One must be sure to ask how well the effects address the need, as well as how important that need is. And one must be cautious in asking how important the effects are, if this questioning is indeed intended to mean how well the effects meet the need and also how important that need is.

(3) The third approach to effectiveness is to use some specified operational standard, such as a test score. This approach would be especially appropriate if all the products in a review could be rated by the same standard. Since the products would not address different needs in this case, no rating for need would be necessary.

If the products do have different standards for effectiveness, however, a need rating becomes essential; otherwise, no discrimination would be made between a product that does well on a simple, narrow standard of effectiveness, and one that does well on a broad, comprehensive standard. The need criterion could be worded: How important is the achievement of this standard of effectiveness?

(4) If the effectiveness question addresses how well the product does what is needed, the need question becomes how important that need is. This approach has the advantage of rating the effectiveness of the product on the basis of what users actually want it to do, and ties together the need and effectiveness criteria most clearly. With the previous three approaches, the possibility always exists that the effectiveness will be rated in terms of objectives, effects, or operationally specified standards that are only vaguely relevant to what is needed, and the fundamental question, how well the product does what we need, will be ignored in the entire review.

(5) The fifth approach is to ask what the product can be used for, and to rate the effectiveness as to how well it can succeed in that capacity.

The need question would be, simply: How important is the use?

E. Summary of approaches to need and effectiveness. The five approaches can be outlined as follows:

(1) The goal definition.

Effectiveness criterion -- how well the product achieves its goals.

Need criteria --

how well the goals address the need;

how important the need is.

(2) The effect definition.

Effectiveness criterion -- how strong an effect the product has.

Need criteria --

how well the effects address the need;

how important the need is.

(3) The operational definition.

Effectiveness criterion -- how well the product does on an operational standard of effectiveness.

Need criterion -- how important the achievement of this standard of effectiveness is.

(4) The need definition.

Effectiveness criterion -- how well the product does what is needed.

Need criterion -- how important that need is.

(5) The use definition.

Effectiveness criterion -- what the product can be used for, and how well it does the job.

Need criterion -- how important that use is.

IV. Problems with Evidence

A. The evidence problem. In a typical review situation, reviewers are faced with evidence that is grossly inadequate. Some claims will be made with no evidence to back them up, some with evidence that does not actually support them, some with evidence of dubious reliability. The evidence may relate only to narrow and scattered aspects of the quality of the product. A product might have good evidence that it is moderately effective; or it might have moderately convincing evidence that it is very effective. A product might seem to be of very high quality but the evaluation data might be faulty or even nonexistent. Or a product might have been shown beyond doubt to attain its objectives, but only because its objectives are trivial.

Nevertheless, reviewers are faced with a situation in which they must reach a decision, no matter how faulty the evidence. Their only alternative is to take the available evidence into account as best they can, giving appropriate weight to reliable and relevant evidence, ignoring unreliable and irrelevant evidence, and keeping in perspective just what has been shown and what has not been shown about the product. The possible ways to do this, and the issues that must be kept in mind, are the subject of this section.

The difficulty of dealing with inadequate and scattered evidence is aggravated by the fact that evidence is potentially relevant to all the criteria, and has a different significance for each one. In many cases the evidence serves to increase our confidence in the reviewer's judgment. If a reviewer gives a textbook a rating of "very good" on social fairness, for example, our confidence in that rating is increased if the producers have provided evidence that outside experts have come to the same conclusion. In other cases the evidence is the

actual information on which the reviewers must base their judgment. This is especially true of evidence for effectiveness; the reviewer can only know what the product actually does from what the product has been shown to do. Here evidence is the actual basis of the judgment, not just a reason for extra confidence in the judgment.

These two different functions of evidence require different treatment in a review. When evidence increases our confidence in a rating, we might want to increase that rating by half a point or a point, or better, we might want to weigh that rating more heavily in the final, overall rating. When evidence is the basis of the judgment, of effectiveness, for example, it seems more appropriate to divide the rating into several parts. First, one could rate the apparent effectiveness, based on the assumption that the evidence is credible. Next, one could rate the credibility of the evidence. Finally, an overall rating of effectiveness would be based on the first two. Perhaps the credibility could be rated as a percentage and multiplied with the apparent effectiveness to get the overall rating.

With each individual criterion we must ask how the evidence that bears on that criterion should be handled. Furthermore, each separate effect, or each separate objective, will have its own evidence that must be considered. Giving one overall rating on the evidence for effectiveness has the disadvantage of tending to ignore how many of the effects are demonstrated and how many are not. For other aspects of the need-effectiveness question, the evidence is less crucial. A reviewer's judgment by itself carries some weight as to whether the goals of a product address the need, and on the importance of that need. With these criteria, therefore, unlike with the criteria used in judging whether the product attains its goals, the evidence can be considered a reason for further confidence in the judgment, not the basis of the judgment.

B. Possible approaches to evidence. We have found three general approaches to the problem of how to incorporate the available evidence into the decision-making process of a review. One is to rate the evidence separately for each criterion, as in the "NIE Product Rating Form" (13). A second is to give a separate rating for the evidence in an extra section of the review, as in the rating form designed by Scriven for the NIE "Summary Report on the Product Quality Review Associated with the 1974 Dissemination Request for Proposals" (14). A third combines aspects of both of these, and is used in the model rating form proposed in Section VII of this paper.

(1) The first approach, rating the evidence separately for each criterion, and perhaps for each effect or objective in the effectiveness section, has the advantage of being the most thorough and of providing the most feedback to the producers and to other reviewers on a panel. It also points out distinctly just what the evidence bears on and what it does not bear on; this is important in the usual situation where the evidence is narrow and rather scattered.

The problem with the first approach is that it gets complicated, both for the reviewer, who has at least twice as many ratings to make, and for whoever has the task of pulling together the many ratings into one final rating, a task that is always difficult. The rating for each criterion must be combined with the rating of its evidence; a combination must be different for different criteria, and even for different kinds of evidence that might be provided for any one criterion. Perhaps some general method of treating the evidence for all criteria could be found, such as raising the rating on a criterion a point or half a point when it is supported by strong evidence, but we have not found one yet. Or perhaps a formula could be devised of increasing the weighting a rating has in reaching the final rating according to the credibility of the evidence

for that rating.

(2) The second approach is to rate evidence separately, as a special section of a review, along with need, effectiveness, practicality, and intrinsic quality. This approach makes it far easier to reach a final overall rating and also to decide exactly how much weighting to give to the rating of the evidence. Also, the reviewer is allowed to decide directly what importance to give the evidence. This is more in keeping with a policy of using reviewers' judgment as fully and freely as possible, rather than using arithmetic to pull together various ratings. It is somewhat illogical, however, to treat evidence as a separate aspect of the product to be rated; evidence is the basis for, or support for, the other aspects, not a separate aspect. But logic could be sacrificed to practicality, if need be.

When the evidence is rated separately, five areas should be considered. These will be discussed in detail in the next section, but briefly they are: First, the quality of the evidence -- whether it is empirical or simply judgmental and anecdotal -- and if empirical, such factors as the experimental design, the use of controls, objectivity, and demonstration of causality should be considered. Second, the issue of the ability to generalize the evidence, the question of whether the product has been shown likely to have the same results in situations other than those of the field tests. Third, the presence of comparative evidence, evidence that can be used to compare the product to other alternatives or critical competitors. Fourth, in order to avoid discrimination against products whose objectives are difficult to measure, creative writing materials as opposed to grammar and spelling, for example, the difficulty of obtaining evidence must be considered. Finally, the completeness and relevance of the evidence must be

evaluated in order to account for evidence that is very convincing but may, in fact, be narrow, scattered, incomplete, or only mildly relevant to what it attempts to demonstrate. This fifth aspect, though usually missed in most reviews, is essential in this approach, which, in contrast to the previous one, does not account directly for what the evidence relates to and what it misses.

(3) The third approach takes into account the special importance of the evidence for effectiveness. The effectiveness section of the review would have three parts. The first part is a rating of apparent effectiveness, made under the assumption that the evidence does support what it intends to demonstrate. The second part is a rating of the credibility of the evidence for effectiveness, based on the five aspects mentioned in the second approach. The third part, the overall rating of effectiveness, would reflect the first two parts. Evidence for all other criteria would be brought together in another section of the review. It is this approach that is used in the model rating form proposed in Section VII.

Since this approach uses only two ratings of evidence, it has most of the advantages of the second approach. Since it singles out the evidence for effectiveness -- the evidence which is the most crucial -- it has part of the advantage of the first approach in treating the evidence according to the criteria it is relevant to. Because the overall rating for effectiveness reflects both the claims made and the evidence for the claims, the rating will account for such possibilities as good products with weak evidence, products with good evidence that indicates they are moderately effective, products with evidence that is very convincing but is limited relative to the claims made, and products with strong evidence that shows that they meet their objectives but only because their objectives are very weak.

One of the difficult cases that we keep mentioning is one which arises when a product's goals are not what is suitable for the need that the product was designed to address. We have suggested earlier that in this case one would want a rating of how well the product attains objectives that <u>are</u> suitable. But now the evidence for effectiveness, no matter how convincing, is not likely to be very relevant to the rating for effectiveness in terms of the desired goals. The rating for the relevance would have to be very low in that case, and also for the overall credibility of the evidence. But this is as it should be; if producers submit evidence that does not support the intended claims, that evidence is worthless, no matter how "scientific" it seems to be.

C. <u>The problem of bias</u>. Good evidence to support a product is extremely difficult to obtain. For one thing, finding appropriate and reliable measures is not always possible. For another, really complete evidence is very expensive to obtain. Long-range studies must be made, alternative products and approaches must be considered for comparative data, adequate control groups must be used, and broad goals requiring many kinds of tests and measures must be examined. It is not surprising, in this light, that producers so often resort to instances of user satisfaction as their only supportive evidence.

The problem of bias that we are concerned with here is this: Some goals of products are easier to get evidence for than others, and products with these goals will be given an advantage in a review that places a heavy emphasis on the quality of the evidence. The designers of some reviews may wish to give this advantage to them, if, for example, they believe that all products should be designed with easily measurable intents. It is suggested that this kind of bias not be built into a review carelessly and unintentionally by simply

emphasizing the quality of the evidence and not considering the result of that emphasis. The bias would tend to favor products that are traditional and that have narrow, straightforward, cognitive objectives. Products that are meant to encourage creativity, curiosity, or student initiative, for example, would be at a great disadvantage.

This bias can be countered easily, if desired, by taking into account the difficulty of obtaining the evidence for each product when rating the quality of the evidence, either by lowering the standards in difficult cases or by giving extra points, in effect, to the rating of the evidence in these cases.

Some reviews actually place so much emphasis on the quality of the evidence that they ignore what it is that the evidence is showing. In "Sharing Educational Success -- A Handbook for Validation of Educational Practices," for example, a program is given the same rating if there is good evidence that it is frightfully expensive as it is if there is good evidence that it is inexpensive, since only the quality of the evidence is considered (17, p. 87). This example is extreme, of course, but there is a definite tendency to distinguish between the policy or value question of setting objectives or criteria and the scientific question of determining whether those objectives or criteria are met, and then to want to focus on the scientific question in the review. At best, this kind of thinking leads to the unintentional bias mentioned above. Sometimes it leads producers to simply state as objectives whatever the evidence seems to support, if they know that the quality of the evidence will have more weight than the importance of the objectives. Often the completeness and relevance of the evidence are ignored, even when the importance of the objectives are given appropriate weightings.

The problems of this extreme emphasis on evidence can be avoided if included in the review is a rating of what the evidence intends to show (our model rating form, for example, rates the apparent effectiveness), and also if the rating on the quality of the evidence includes consideration of the completeness and relevance of that evidence.

D. The problem of relevance. With some kinds of criteria, and with some kinds of objectives, the supportive evidence either shows that the criteria or objectives are met or it does not show it. A criterion that considers if the pronoun "he" is never used to mean "he or she," or an objective that specifies the amount an average score will be raised on a specified test can be supported by clear-cut evidence. Operational objectives, in general, will be like this. Rating the evidence for these objectives or for these criteria is a relatively simple matter of looking for flaws in the evidence -- inadequate samples, controls, or measures, for example.

But evidence that is meant to justify, for instance, a rating of 3 on the criterion of sex-role fairness or on the criterion of strength of effects requires extra consideration. Supposing the evidence is conclusive that test scores have been raised a certain amount: Is that supportive of a rating of 3 or only of 2½? This is the question of the relevance of the evidence, which could be worded, given that the evidence shows what it purports to show: How relevant is it to the rating which you are giving on the criterion in question?

Any rating of the evidence for criteria that are scaled must include consideration of relevance. Otherwise it is very possible that evidence that

clearly shows one thing will be used to justify something else. There is no simple way to distinguish between cases where relevance is important and those where it is not. When questioning whether the product meets its objectives, relevance does not come up; but in questioning how well the product meets its objectives, relevance is important, whether those objectives are operationally specified or not. Even screening criteria sometimes raise this issue, for screens often imply a rating of "passable," and in that case, one must ask if the evidence, even if factually correct, actually justifies such a rating.

Single purpose reviews, especially those in which all the products have the same objectives, may not have this relevance problem, if all the products have the same kind of evidence and the evidence is therefore equally relevant in each case. Perhaps a group of reading products all would have evidence based on the increase in score on the same reading test. Here, only the quality of the evidence needs to be rated -- the adequacy of controls, samples, statistical analysis, and so on. Even in this rather straightforward case, however, complications can easily arise: How can one be sure that the increase in test scores is not based simply on improved vocabulary or improved test taking ability, unless evidence is provided to show that these were not stressed in the product? And with evidence like that consideration of its relevance is again significant to be sure that the evidence shows what needs to be shown.

In short, examining the relevance of the evidence is essential in distinguishing between evidence that a product is good and good evidence. Data can be good evidence in the sense of proving something conclusively, but still not be evidence that the product is a good one if what is proven is not what needs to be proven. Only when the relevance of the evidence is considered can one be sure that one has good evidence that a product is a good or effective one.

V. Criteria

In this section each criterion that may be included in a review is discussed. The list is meant to be exhaustive; designers of reviews could go through this list to be sure that their reviews include all the criteria that are relevant in their situation. It could be thought of as a menu of possible criteria. In an effort to be exhaustive some criteria are included which will not be relevant to all types of reviews. Also, a few of the criteria are somewhat redundant, repeated with somewhat different perspectives in cases where these different perspectives could not be included in one general criterion, or where one perspective required the criterion to be placed in one section of the review and the other required its placement another section. No review, therefore, would be expected to have all 29 of these criteria. The model rating form, in Section VII of this report, shows one selection and grouping of criteria. An outline of the criteria is presented on the next page.

List of Criteria

Need

1. Priority of the need addressed
2. How well the goals address the need
3. Appropriateness to users
4. Cumulative effects
5. Product uniqueness
6. Social acceptability
7. Marketability

Intrinsic quality

8. Social fairness
    a. sex
    b. race
    c. ethnicity
    d. religion
    e. socioeconomic status
    f. age
9. Choice and accuracy of contents
10. Technical quality
11. Instructional quality
12. Product appeal and learner satisfaction
13. Adequacy of development process

Practicality

14. Costs -- initial, ongoing, and competitive
15. Staff effort needed
16. Acceptability to users
17. Adoptability -- exportability, flexibility, and manageability

Effectiveness

18. Strength of effect
19. Long term effects
20. Comprehensiveness or number of effects
21. Scope of the effect
22. Comparative effectiveness
23. Efficiency in use of learners' time
24. Side effects

Credibility of the evidence

25. Relevance and completeness
26. Quality of the evidence
27. Extent to which evidence can be generalized
28. Comparative evidence
29. Difficulty of obtaining evidence

A. Need criteria. Need criteria rate the need for the goals or effects of a product, rate the importance of the need addressed by the product, or rate other reasons for giving an especially high or low priority to a product. The specific criteria are:

(1) Priority of the need addressed. This includes consideration of the social and educational significance of the need or problem addressed by the product, the size of the target population, and the priorities of the reviewing agency. This criterion is especially important when the allocation of funds depends on the outcome of the review. High priority products, in this case, are most important. A review designed to aid in selecting products to include in a catalog, on the other hand, could include high and low priority products and leave to users the decision of priorities.

The wording of this criterion in any review must be clear about several issues. Is the need being rated the one addressed by the producers? If so, is it possible that the product could be used to address some other need that is more urgent? And if there is a more urgent need, where is it being rated? Can the product be used to address several different needs? Is the rating supposed to reflect the most urgent needs or the aggregate of all needs? If the product does not address the need that it is supposed to, where can the reviewer rate the importance of the need that it does address?

Perhaps rating this criterion is most difficult when choosing between a product that addresses an intense need for a small target group -- a special education product, for example -- and one that relates to a less crucial need shared by many students. But no wording of the criterion is going to make easier this fundamentally difficult question of priority.

(2) How well the goals address the need. That is, is it directly relevant to the need, is it central to the need? This is the second part of the general needs-effectiveness question discussed in Section III. This criterion could be combined with the first, if desired, and then would be worded: What is the priority of the goals? If effectiveness is rated in terms of strength of effects rather than in terms of goals, the question could be reworded: How well do the effects address the need? Or, in combination with (1), what is the priority of the effects?

In cases where the goals do not address the need very well, one will want to know how well the product attains goals that would address the need, now how well it attains its irrelevant goals; provision should be made for discovering this in the review. Our model rating form approaches this issue directly by first asking the reviewers to specify goals that do address the need, and ratings are done in terms of those goals. Such an approach makes this criterion redundant in a sense; it does no good to ask reviewers to rate the relevance of goals that they have suggested themselves, since they of course will think them appropriate. Accordingly, this criterion is omitted from the model rating form.

(3) Appropriateness to users. Does the product "fit" with current, ongoing programs? Does it meet needs unmet by other products currently in use? This could be called the content question: Do the contents cover what needs to be covered? For many user reviews, this is a crucial question. What users need is a product to teach certain skills, subject areas, or attitudes. The need for the product depends on whether it covers those areas. In a national-level review, however, this criterion cannot be handled, except in cases where a certain content area is known to be lacking.

(4) Cumulative effects. Some products have effects that are cumulative, and may be given extra priority as a result. Examples are products to teach people to teach others, such as teacher-training products, or products that teach very broad, general, or fundamental skills, such as reading, that are useful in later education. Any product that has a secondary effect that occurs after its initial effect should be considered from this perspective.

(5) Product uniqueness. This criterion is relevant if the distribution of funds depends on the outcome of the review and if the review is meant to have what we call a research function. In that case, the agency may wish to give extra priority to a product that is unique, in order to try a new approach. Uniqueness is included under the need criteria because it is a matter of priority in some cases; however, the fact that a product is unique is not necessarily a reason to believe it is of high quality.

(6) Social acceptability. We do not know quite what to do with this criterion, except to mention it as a possible one that could be included in some reviews. It would include such issues as the social acceptability to the communities with which the product will be used, or whether it meets any legal requirements. Some aspects of this question seem perfectly legitimate. For example, some communities may want an emphasis on traditional education, others on freer and more open approaches, some on liberal education, others on career training. These are value issues that a community should be free to decide on. The desire for educational material that promotes patriotism, views of morality or religion or censorship of viewpoints are criteria of a questionable nature. But it is hard, in many cases, to distinguish legitimate issues from illegitimate ones, and each group conducting a review must decide how it wants to deal with this criterion.

It must be pointed out that sex and race fairness are also moral issues, and ones in which it is national policy to promote a certain viewpoint. If it were not for the policies on these issues, they too would have to be covered in this criterion of social acceptability. As they are national policies, they are considered separately.

(7) Marketability. This is only an issue in certain reviews, designed to determine dissemination funds and packaging of programs. Users choosing a product would have no concern for whether the product is marketable or not, only whether it meets their needs. This criterion includes several disparate and sometimes almost contradictory elements, depending on the purpose of the review. These are, whether the product would be used if available, whether it actually needs support, and whether support would be helpful. A highly marketable product may not need support; an unmarketable one might not be helped even with support.

B. Intrinsic quality. This is a broad category of criteria and includes many that have little in common with one other. Some are what can be called secondary indicators, that is, indirect indications that the product might be especially effective. In general, all they have in common is that the ratings of the product on these criteria are usually formed by the reviewer's judgment based on an examination of the material itself.

(8) Social fairness. This is a rating of the extent to which a product includes or excludes bias on the basis of sex, race, ethnicity, religion, socio-economic status, or age. Specifying how to make this judgment is difficult. To start with, each reviewing agency must decide on a policy to deal with products that present an uncritical portrayal of the status quo. Is this acceptable, or

must products always be sure to point out unfair situations or the result of these unfair situations when they come up? Also, the agency must have a policy on the extent to which bias in products can be allowed if some provision can be made for countering it in the classroom. This is especially an issue with some forms of literature, for example, which are important in spite of racist or sexist content.

Social fairness, like many of the intrinsic quality criteria, may be used as a screen in an initial screening stage of a review. Another consideration is to rate the product for social fairness and either eliminate the product (or send it back for modifications), or include the rating in the overall rating. If social fairness is used only as a screen, and not part of the overall rating, then the final rating will not reflect whether the product is merely passable or is outstanding in this criterion. For some products, a chemistry workbook, for instance, this may not be of great importance. On the other hand, in a career education textbook or film, one would certainly want to give preference to materials that give a fair treatment to all groups, and not simply screen out the unacceptable ones.

Reviewers may find it easier to make judgments on social fairness if the criterion is broken down into many smaller criteria. Appendix C suggests references to materials that have been developed to do just this. For example, race and ethnic fairness can be broken down into these criteria suggested by the NEA for history texts (18):

a. Does the material promote mutual understanding and respect between races and ethnic groups?

b. Does the material stress the value of cultural pluralism?

c. Does it include the contributions of various groups?

d. Does it emphasize the equal worthiness of all groups?

e. Does it point out the historical disadvantages of various groups?

f. Does it point out the current disadvantages facing each group?

Sex fairness might be broken down as follows (suggested by McGraw-Hill)(15):

a. Does the material avoid stereotyped roles, especially in jobs?

b. Are people treated primarily as individuals, rather than as members of one sex?

c. Are there examples of working mothers, as one possibility?

d. Is the use of the pronoun "he" to mean "he or she" avoided?

e. Are both men and women shown in all kinds of home activities?

f. Are boys and girls both shown as having a range of school interests?

g. Are men and women used equally as examples and in pictures?

Such breakdowns are helpful in making the social fairness criteria operational.

(9) Choice and accuracy of contents. This criterion is straightforward and requires little explanation, though it is of course very important. One would not want to use a product that is very effective in teaching inaccurate material. The choice of contents should be appropriate both to the subject matter and to the level of the learners, although experts may of course disagree on what content is important and what superfluous. Accuracy too depends on the level of the learners. The degree of accuracy appropriate in a sixth-grade science film would not be adequate for an advanced physics class.

This criterion overlaps somewhat with appropriateness to users. But appropriateness to users is a need criterion, dealing with the question: What is the need for the product, inclusive of contents as well as other criteria, in the actual situation where it will be used? The intrinsic quality criterion is: Are the

contents appropriate for what the product is trying to accomplish?

This criterion, like the previous one, could be used as a screen or as both a screen and a rating in a review. One would want to reject a product whose contents are inaccurate, regardless of its rating on other criteria.

(10) Technical quality. Is the presentation clear and economical? Is the language appropriate, proper, and grammatical? Are the printing, photographs, and graphics of high quality? Is the choice of medium reasonable? Is the product well made and durable? Is it well organized and edited?

(11) Instructional quality. Does the product use a sensible teaching strategy? Is the teaching strategy appropriate for the target group? Are the reading level, pacing, and sequencing appropriate? Is the methodology clear and suitable? Is the product internally consistent? Are the evaluation components helpful, diagnostic, well designed?

This criterion can be viewed as a secondary indicator of effectiveness, though there is little evidence to indicate that products which seem to be of high instructional quality will in fact be more effective. A much more important significance of this criterion is that many of the values students learn about education, learning, curiosity, and authority are acquired because of the instructional strategy used by their textbooks and other materials. In many ways, how students learn may be as important as what they learn.

(12) Product appeal and learner satisfaction. Is the product interesting and physically appealing? Is it condescending or just plain boring? Like the previous criterion, this one is of dubious value as a secondary indicator of

effectiveness, and gains its importance from another factor. It seems plausible at least that students who find education fascinating will be more likely to have a favorable attitude toward learning and to continue their education, either formally or otherwise, more willingly. Appealing and satisfying products can help achieve this end. These products are also more likely to be chosen and to be used by learners than other products.

(13) Adequacy of development process. This criterion should go beyond simply asking who the producer is. It includes whether the product was developed using a well designed, scientific process? Were external evaluations used? Were the goals and objectives explicit? Was the development process guided by the goals and objectives?

One important aspect of the development process is whether the product was learner verified and revised, and if so, whether it was done thoroughly, systematically, and empirically. One may also ask whether continuing revision is planned for the entire life of the product. The use of L-BR could be used as a screen, and must be in those states where it is required.

C. Practicality.

(14) Costs -- initial, ongoing, and competitive. This includes the cost of the product, replacement of consumables, installation costs, personnel costs both initially and ongoing, and material and equipment costs. Comparative costs, considering both competitive products and common alternatives (such as teachers' lectures) must be included. Generally, the only way to rate the cost is comparatively since it is the only way to have standards for rating. That is why comparative costs have been included in an overall cost rating.

(15) Staff effort needed. Does the product require extra time and energy from the staff, either professional or support staff, in training or in day-to-day preparation? This criterion could be given a monetary value and then be included in the previous criterion as another personnel cost, but it may tend to be ignored.

(16) Acceptability to users. Will the product face opposition from administrators, teachers or students? Are new methods of teaching or learning likely to cause resistance? Is controversial subject matter introduced?

(17) Adoptability. This can be broken into several aspects. (a) Exportability: Is the product actually usable in sites other than the development site? Can it stand alone? If special training or equipment is needed, are they available? (b) Flexibility: Can the product be adapted to existing constraints, equipment, and personnel? Can it be adapted to local needs? (c) Manageability: Does the product create problems of management? Are organizational changes needed in scheduling, product maintenance, or use of resources?

These three aspects are separate in theory but would tend to be hard to distinguish in practice. For example, is a product that requires unusually well trained teachers to be labeled not exportable, inflexible, or unmanageable? Because these aspects do overlap, we have combined them into one criterion.

D. Effectiveness. In Section IV of this report it is pointed out that there are several methods of handling ratings for effectiveness and the evidence for those ratings. The method chosen for our model rating form involves first rating apparent effectiveness and the credibility of the evidence, and then reaching a final, overall rating for effectiveness from those two. By this method, the effectiveness criteria we are about to discuss would be included

under apparent effectiveness. No matter what method is used, however, these same criteria must be covered.

(18) Strength of effect. This is the overall rating for how powerful an effect the product has. It could be rated in terms of how well the product meets its goals or objectives, or also in terms of how well it attains goals or objectives that the reviewer would expect from it. Also, each goal or each effect could be rated independently in some cases.

If rated simply as strength of effect, it would be helpful to ask reviewers to be explicit about what standards they are using. If rated in terms of its own objectives or goals, provision should be made for these two possibilities: First, that the product might not attain its own objectives very well, but might attain other objectives equally important. Actually, this is unlikely because of a tendency of producers to list any known effect of a product as an objective. Second, and more likely, that the stated objectives are not actually the ones the reviewers would be interested in, so that the important rating of effectiveness would be whether the product attains objectives the reviewer thinks it should have.

(19) Long term effects. These are harder to demonstrate but are of course much more important when they are demonstrated. Skills and fundamental subjects that are remembered are more significant than knowledge that might be forgotten. Included with long term effects would be impact effects. These are the effects on pupils whose teachers are using a certain product, usually a teacher training product. Impact effects are, after all, the ultimate purpose of teacher training.

(20) Comprehensiveness or number of effects. Knowledge, skills, attitude, and awareness changes should all be considered. The rating could reflect what is

called for by the needs addressed so that a product that is designed to meet a narrow, very specific need would not get a lower rating as long as its effects were as comprehensive as called for by that specific need. When rated by this standard, the criterion is equivalent to the second criterion: the relevance of the effects or goals to the need. On the other hand, if this criterion is simply thought of as the number of effects so that a product with broad goals can achieve a higher rating than one with a narrow goal, there is a danger of overlapping the first criterion which normally includes the importance of the goals. Whichever way this criterion is handled, it must be unambiguous to the reviewers, for they may well give very different ratings to those products with narrow goals.

(21) Scope of the effect. How widespread is the effect, in terms of the percentage of the target group that is affected, as well as nontarget users? Some products may be effective only with a portion of the users, such as the best readers, or the most mathematically inclined. Together with the two preceding ones, this criterion makes up what could be considered the three dimensions of effects: how strong they are, how many there are, and how many students are affected. These dimensions can vary independently and must all be considered.

(22) Comparative effectiveness. If the previous criteria were not rated using a comparative standard, then comparative effectiveness must certainly be included as another criterion. Comparison could be either to competitive products or to alternative methods of accomplishing the same goals.

(23) Efficiency in the use of learners' time. Do other methods or products accomplish the same result in less time? This criterion has been used in industrial training programs where the employer is paying for the time lost by the workers.

In schools, little attention has been paid to it. This criterion is different for the efficiency in the use of teacher time, which is considered under practicality.

(24) Side effects. In any review rating effectiveness in terms of the goals or objectives of a product, the possibility arises of powerful side effects. This would include interactive effects, that is, ways that the product interacts with other materials or methods to either supplement or counteract them. Also, any negative effects would be included here as well as any assurance the producer makes, or neglects to make, about harmlessness.

E. Credibility of the evidence. Whatever the method of incorporating a rating on the credibility of evidence into the final rating, these criteria should be considered:

(25) Relevance and completeness. Does the evidence fully justify the rating that it is based on? Is it incomplete? Does it leave a gap between what is claimed and the evidence that is supposed to support that claim?

(26) Quality of the evidence. Is the evidence empirical or just judgmental? If empirical, some of the considerations that arise are:

(a) the size and choice of the sample

(b) suitability and implementation of the experimental design

(c) the use of controls

(d) the use of measures (appropriate, reliable, and valid)

(e) quality of statistical analysis

(f) demonstration of causality, including consideration of alternate explanations

(27) Extent to which evidence can be generalized. Even if evidence shows conclusively that a product has a certain effect in a test situation, we must consider the possibility that it may not have a similar effect in other situations. Perhaps the effect would only occur in exceptional circumstances, with a teacher who is outstanding, for example, or perhaps one who has a particular teaching style. So one must consider if and to what extent the evidence can be generalized to other situations.

(28) Comparative evidence. Does the evidence include data on competitive products or alternative methods of achieving similar goals? It would be difficult for evidence to justify a high rating without including this.

(29) Difficulty of obtaining evidence. Since some products will be harder to obtain evidence for than others, this difficulty must be rated in order to avoid giving an advantage to products with easily proved objectives, unless the reviewing group wishes to give this advantage. Objectives such as teaching creative writing, or increasing student initiative, curiosity, or promoting attitude changes in general, will be harder to demonstrate than traditional, cognitive objectives.

VI. Practicalities of Reviews

We would like to make a few points about the practicalities of carrying out a review:

A. The review process can be simplified by beginning with a screening stage or stages. This is usually necessary to narrow down the number of products to

those with the appropriate subject matter. The more narrow the subject, the easier the review process. Rating and choosing between very diverse products precludes the possibility of having clear and evident standards by which to make judgments. Sometimes it is possible to divide a review into several separate reviews, each with its own subject area, eliminating the need to do actual comparisons between very different products. If a screening stage of a review can eliminate, or separate into categories, products that have different goals, or that address different needs, the whole need section of the review may be made unnecessary.

B. A verification stage may precede the actual rating of products. This need not be done by the experts who form the review panel, and assures the panel of receiving information that is somewhat reliable. The verification stage is much more useful and also simpler if producers have been encouraged to make specific claims and assurances about their products, and to be specific about the evidence that supports these claims and assurances. Assurances are usually made about screening criteria, with which products can be expected to meet some minimum standard, such as harmlessness, social fairness, transportability, content accuracy, and adequacy of the developmental process. Claims could be made concerning any criteria, including the same ones about which assurances are made. An assurance that the contents are accurate could be supplemented by a claim that they include the latest findings.

The claims and assurances can be verified by the staff of the reviewing agency prior to the actual review. Rating the significance adequacy of the claims and assurances, however, is not a verification activity; it is a review activity, and should not be disguised as part of the process of verification. If products are to be screened for some level of adequacy of the claims and

and assurances, this must be considered a review activity, and conducted by appropriate personnel.

It is doubtful that it will be common to conduct an entire review by simply rating the significance of each claim and assurance and the evidence supporting them as opposed to having set criteria by which products are rated. Claims and assurances tend to be very scattered and incomplete. A review that considers claims and assurances only would probably omit too many important criteria. Perhaps producers could be convinced to submit claims and assurances about a set range of criteria, however.

C. Reviews are most commonly conducted by small panels, large panels are too expensive. Reviews done by one individual or even a group of individuals whose ratings are simply averaged, lose the important opportunity to discuss the judgments that are involved. A calibration session before the actual review helps assure that the panel members interpret the various criteria consistently and have some idea of common standards. If each product is rated by the panel members individually before the panel meets, the voices of the less outspoken members are more likely to be heard, and more issues are likely to be raised, than if the panel does the rating as a group from the outset. When the individual members are making their ratings, they might be encouraged to be explicit about their frame of reference, assumptions, and underlying values. Then these factors can be dealt with by the panel when they are relevant. The panel should represent as wide a range of interests as feasible, for each viewpoint has something to contribute: Experts in the content area and in evaluation, teachers, administrators, students, consumer groups, and interested community members can all be included.

D. The main purpose of having set criteria by which to judge products is

to be sure that every issue is covered, thoroughly and systematically. The criteria can be thought of as a checklist of things to consider. Different ones will be important for different products, but at least one can be sure that no important issue is omitted. This purpose is probably more important than trying to use criteria to get a set of numerical ratings which, when averaged with appropriate weightings, will give an overall rating. Perhaps a set of criteria which can be used in this way for all products will some day be developed, but we have found no sign of it at present.

Our view of the use of criteria and rating forms in general is well expressed by Epstein et al. in their report of the ETS review (24, p. 42):

> The panel also emphasized in its reprise that the rating forms turned out to be outstandingly useful tools to help focus their review of products upon issues of evaluation. The forms proved an ideal way for panelists to collect their judgments systematically and "store" them for later referral. Their more important strength, though, was that they facilitated the move from individual to group judgments about the product at hand.
>
> ...In a sense, then, the rating form may best be regarded as a paper upon which the panelist wrote a memo to himself [or herself] .

E. The final, overall rating for each product must be made by the panel. No simple mathematical formula is likely to reach a convincing final rating from weighted averages of the ratings on the various criteria; criteria would certainly need different weightings for different kinds of products, and possibly even for different products with common goals. The fewer the number of criteria that are rated, the easier it is to reach a final rating. If the main concern is to reach a final decision, grouping criteria to be rated together is useful. To get maximum feedback to producers, however, the greater the number of ratings the more detailed the information.

We believe that the results of a review are only as good as the judgment of

the reviewers. No mechanical method of conducting a review has been found which would allow an inexperienced staff to reach a dependable conclusion about materials in areas they are unfamiliar with. Since the review is so dependent on the reviewers' judgment, the review should be designed to take full advantage of that judgment by letting the reviewers make the final rating of each product.

VII. Model Rating Form

A. Introduction. This section presents a model rating form that takes into the account the various issues raised in this report. We decided the model would be most useful if it is specific. Accordingly, we chose a specific kind of review, namely, a national-level review to decide on the allocation of funds for the dissemination of educational products.

For the same reason we also chose one specific alternative whenever there were several methods of handling an issue or criterion. In particular, the need-effectiveness question is handled by asking the reviewers to analyze the needs and suggest objectives that would be appropriate; the rating of effectiveness is done in terms of these objectives. This approach was chosen because it was felt that the producer's goals and objectives could not be taken for granted. They may be suitable, but they may not be, and we wanted to leave this important decision to the reviewers.

We realize that many people will not agree on this approach. Only small changes are necessary in the rating form to make it consistent with another approach. Some of the most likely modifications will be suggested in part c after presenting the form.

B. The form.

Model Rating Form, p. 1

Model is for a decision to distribute funds for dissemination.

i. Background.

What need and target group do you see this product addressing? If different from the producer's view, be explicit.

What objectives do you think this product should have? Again, be explicit if they differ from the producer's.

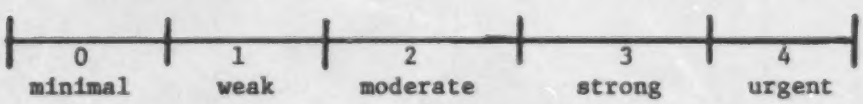
What do you consider common alternatives or critical competitors to this product?

Make explicit any frame of reference, assumptions, or values that are influencing your judgment of this product and which might be important for other panel members to be aware of.

Model Rating Form, p. 2

ii. Need: the need for the product's goals or effects.

Place an X above the appropriate number.



Consider:

___priority of need addressed (i.e., that you think it addresses).

___size of affected population.

___appropriateness to users: Does it fit in with ongoing programs, other materials?

___product uniqueness: Should product be promoted in order to try a new approach?

___cumulative effects -- teaching people to teach others; general skills.

___marketability: Does it need support, will support help?

If any considerations are especially important, either positively or negatively, for your rating, mark them + or -, and if extremely important, ++ or --.

Are there assumptions, values, a frame of reference, or explanations you wish to make explicit for other panel members?

Model Rating Form, p. 3

iii. Intrinsic quality of product.

0	1	2	3	4
poor	weak	passable	good	excellent

Consider:

___social fairness -- excludes or counters bias.

___racial

___sex

___ethnic

___religious

___other (specify)

___choice and accuracy of contents.

___technical quality -- clarity, editing, printing, physical quality.

___instructional quality -- sensible teaching strategy.

___product appeal and learner satisfaction.

___adequacy of development process -- LVR, external reviews, scientific approach.

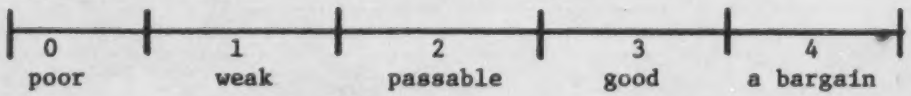
Mark important considerations +, ++, -, or --.

If you think the product should be rejected because of one on these considerations (sex or race fairness, e.g.), mark that with an (*) and explain below. Try to give an overall quality rating, taking that factor as only one of many -- this would be more informative feedback for future revision than an automatic rating of zero.

Explanations:

Model Rating Form, p. 4

iv. Practicality.



Consider:

___costs -- initial, ongoing, and comparative.

___staff effort needed -- training, daily preparation, extra staff.

___acceptability to users -- lack of opposition from teachers, students, administration.

___adoptability, exportability, flexibility, and manageability: can it be adapted to existing constraints?

Mark +, ++, -, or --.

Explanations:

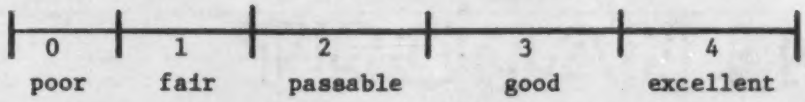
Model Rating Form, p. 5

v. Effectiveness.

The overall rating for effectiveness will be based on two other ratings,

one for apparent effectiveness, the other for the credibility of the evidence.

a. Apparent effectiveness. Accepting for the moment what the evidence purports to show, how effective do you think the product is:



Consider:

___how well product meets objectives you judge to be appropriate.

___long term effects.

___comprehensiveness or number of effects, relative to what is called for to meet the need. Include knowledge, skill, attitude, and awareness changes.

___scope of the effect, in terms of per cent of target group affected.

___comparative effectiveness, relative to critical competitors and common alternatives.

___efficiency in use of learners' time.

___side effects -- effects other than those connected with the objectives you are suggesting; negative effects or harmlessness.

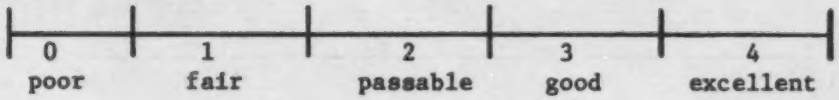
Mark +, ++, -, --.

Explanations:

v. Effectiveness (continued)

b. Credibility of the evidence for effectiveness.

How credible is the evidence in justifying your rating of apparent effectiveness?



Consider:

___relevance and completeness of the evidence: assuming the evidence shows what it purports to show, how well would that justify your numerical rating of apparent effectiveness?

___quality of evidence.

___judgmental or actual empirical evidence? If empirical:

___sample size and choice.

___suitable experimental design.

___adequate controls.

___demonstration of causality.

___objectivity.

___extent to which evidence can be generalized: Will product be just as effective in other situations?

___comparative evidence -- evidence that the product is better than others.

___difficulty of obtaining evidence: Give ample allowance for the difficulty of obtaining evidence for some products, to avoid an unfair advantage to products with simple objectives.

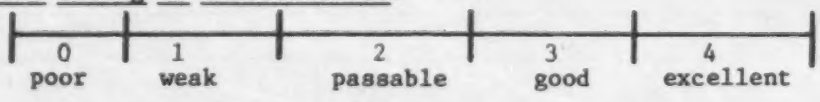
Mark +, ++, -, or --.

Explanations:

Model Rating Form, p. 7

v. Effectiveness (continued)

c. Overall rating of effectiveness.



Consider:

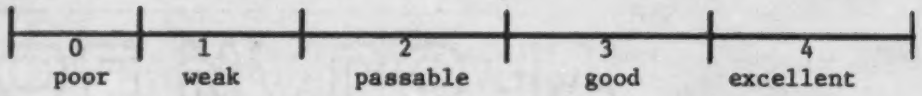
___apparent effectiveness.

___credibility of the evidence for effectiveness.

Mark +, ++, -, or --.

Explanations:

vi. Quality of other evidence.



Consider:

___credibility: Is data judgmental or empirical?

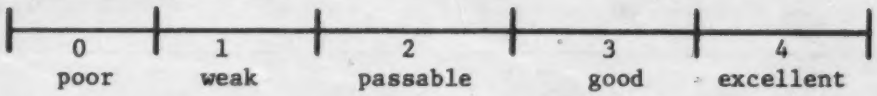
___relevance: Does the evidence show what it is meant to show?

___completeness: Is evidence complete on many criteria or just a few?

Mark +, ++, -, or --.

Explanations:

vii. Overall rating of product.



Consider:

___need.

___intrinsic quality.

___practicality.

___overall rating of effectiveness.

___quality of other evidence.

Mark +, ++, -, or --.

Did you indicate in part iii that the product should be rejected

because of one of the intrinsic quality considerations? ____

Explanations:

C. Possible modifications of the model rating form. The most likely changes that others may want to make in the proposed form are:

(1) In part i, the question of the objectives that the product should have can be omitted. But some method must be used in order to be sure that the goals or effects of the product actually address the needs. Perhaps an additional criterion can be developed. It is unnecessary in the model form since a review based on the objectives the reviewers think are appropriate does not have to ask if the reviewers think those objectives are appropriate.

If this modification is made, some corresponding modification is necessary in the first criterion of part v, section a, which asks for a rating of how well the product meets those objectives which the reviewer has suggested.

(2) In part ii, the first criterion could be designed to ask for a rating of the priority of need that is claimed to be addressed, rather than the need that the reviewer thinks is addressed. Then there would be the possibility that the product does not, in the eyes of the reviewers, address the need that is claimed. In this case, one would want a rating somewhere for the priority of the need that is addressed.

(3) In part v, section a, the first criteria, which deals with the strength of effects, could be handled in other ways, as discussed in Section III of this report. Instead of asking for a rating of how well the product meets objectives which the reviewer deems to be appropriate, it could ask simply for a rating for the strength of the effects. In this case, some standards should be specified, such as in comparison to other similar products, or in terms of what is needed.

This criterion could also be in terms of the producer's goals or objectives. If it seems possible that those goals are inappropriate, a chance should be given to rate other, more appropriate goals. Or if those goals are attained, perhaps others are as well.

However this criterion is handled, the need criteria, especially the question of the priority of the need addressed, should be worded in a corresponding fashion.

(4) Instead of having the criteria grouped together in broad categories, the reviewers could be asked to rate every criterion separately, and perhaps rate the evidence for every criterion.

(5) The evidence could be handled in various other ways, as discussed in Section IV of this paper. The evidence could be rated all together, in effect omitting part v, sections b and c of the form and leaving out part vi. Or the evidence could be rated for each part of the form or even for each criterion.

(6) The summary section, part vii, could be omitted. The final decisions could be made by the reviewing agency based on the ratings of the various reviewers. Or the summary ratings could be simply a weighted average of the previous ratings.

VIII. Sketches of Reported Reviews

In order to facilitate further investigation of review activities, we are including thumbnail sketches of the reviews on which our discussion and conclusions are based. We will point out which reports have been the most useful. The reports that are the most helpful are not necessarily those of the best reviews. Mostly they are the reviews that are the best reported. Many of the

reviews were conducted as part of a larger project, and the project reports may not have included much information about the actual reviewing process.

Our sketches include the following information:

(1) Type of material: program, practice, or product.

(2) Mixed or special topic.

(3) Type of decision to be made: to continue development, to package a program, to disseminate, to use, or to recommend.

(4) Which criteria are used, either as screens or as scaled criteria. The list of possible criteria is on page 39; they are discussed in Section V of this report. We try to be generous in our descriptions of which criteria are considered. If any aspect of a criteria could affect the final rating of a product, we listed that criteria as being included in that review.

(5) Who are the reviewers? The possibilities are the developer's staff, the staff of the review project, outside experts, or users.

(6) Is the product itself reviewed, or just information about it?

(7) Is each product compared with others; are comparative ratings included?

(8) Are rating scales used, or are the criteria just kept in mind?

(9) Is the review done by a panel, an individual, or a collection of individuals with no collaboration?

(10) Is empirical evaluation data required, sometimes present, collected as part of the review project, or not used?

A. Promising Educational Programs and Practices

Reported by Automated Services, Inc. (1).

1. programs and practices

2. mixed topic

3. decision to recommend

4. screening criteria:

   no. 14 -- costs

   no. 17 -- adoptability

   no. 18 -- strength of effect

5. done by staff of review project
6. just information was reviewed
7. no comparisons are made
8. does not use scales
9. panel
10. evidence sometimes present

The purpose of this review is for the preparation of a catalog of promising programs and practices. Few details given on the review process.

B. Reading Programs

Reported by American Institutes for Research (2).

1. programs
2. special topic
3. to package and to recommend
4. scaled criteria:

   no. 14 -- costs

   no. 17 -- adoptability

   no. 18 -- strength of effect

   no. 22 -- comparative effectiveness

   no. 25 -- relevance of evidence

   no. 26 -- quality of evidence

   no. 28 -- comparative evidence

5. staff of review project
6. just information is reviewed
7. comparisons are made
8. uses scales
9. panel
10. evidence required

A report on procedures used in selecting and packaging reading programs. Full information on rating procedure, and copies of the forms used, are included. Selection of criteria is dictated by the request for proposals and not discussed. Rating, which is done by computer on the basis of standardized information from program directors, tends to be very mechanical.

C. Protocol Materials #1

Reported by Development and Evaluation Associates (3).

1. products
2. special topic
3. to disseminate
4. scaled criteria

no. 1 -- priority of need	no. 20 -- comprehensiveness
no. 2 -- do goals address need?	no. 22 -- comparative effects
no. 3 -- appropriateness	no. 25 -- relevance of evidence
no. 9 -- accuracy of contents	no. 26 -- quality of evidence
no. 10 -- technical quality	
no. 11 -- instructional quality	
no. 13 -- development process	

5. experts
6. product and information are reviewed
7. comparisons are made
8. uses scales
9. collection of individuals
10. evidence sometimes available

A report on a conference in which protocol materials are reviewed and discussed. Report focuses on the results of the review rather than the procedures or criteria used.

D. Effectiveness of Educational Laboratories and Centers

Reported by the American Institute for Research (4).

1. products
2. mixed topic
3. to recommend -- but not actual decision
4. scaled criteria:

no. 1 -- priority of need	no. 10 -- technical quality
no. 2 -- do goals address need?	no. 14 -- costs
no. 7 -- marketability	no. 18 -- strength of effect
no. 9 -- accuracy of contents	no. 20 -- comprehensiveness

5. experts
6. product itself is reviewed
7. no comparisons are made
8. uses scales
9. panel
10. evidence sometimes available

Reports the development and trial use of a reviewing procedure, for both instructional and knowledge products. This report discusses every aspect of designing and testing a review. Only a few criteria are used, however, with no rating of evidence. The results of the trial use of the reviewing procedure is presented.

E. Training Materials

Reported by the National Center for Educational Research and Development (9).

1. programs, products, and practices
2. special topic

3. catalog review

4. scaled criteria:

no. 1 -- priority of need	no. 11 -- instructional quality
no. 2 -- do goals address need?	no. 14 -- costs
no. 9 -- accuracy of contents	no. 18 -- strength of effect
no. 10 -- technical quality	

5. staff of review project

6. information only reviewed

7. no comparisons are made

8. does not use scales

9. panel

10. evidence not used

Few details given on review procedure.

F. Protocol Materials #2

Reported by the National Center for the Improvement of Educational Systems (10).

1. products

2. special topic

3. to continue development

4. scaled criteria:

no. 1 -- priority of need	no. 20 -- comprehensiveness
no. 2 -- do goals address need?	no. 21 -- scope of effect
no. 3 - appropriateness	no. 23 -- efficiency to learners
no. 12 -- product appeal	no. 26 -- quality of evidence
no. 18 -- strength of effect	

5. developer's staff
6. product itself reviewed
7. no comparisons are made
8. does not use scales
9. collection of individuals
10. evidence required

Not actually a review, this report proposes a method that could be used in reviewing protocol materials and is included for that reason.

G. Mini-Panel Review of Teacher Training Products
Reported by NIE (12).

1. products
2. special topic
3. to continue development
4. scaled criteria:
only no. 29 -- difficulty of obtaining evidence, was not included.
5. experts
6. information and products are reviewed
7. comparisons are made
8. uses scales
9. panel
10. evidence required

One of the most interesting discussions of issues that comes up in designing review activities and in choosing the criteria. Complete description of the procedure followed and its advantages, disadvantages, costs, as well as the

ratings received by each product for each criterion. Used NIE Product Rating Form (13).

H. Dissemination Review

Reported by NIE (14).

1. products
2. mixed topic
3. to disseminate
4. scaled criteria:

no. 1 -- priority of need	no. 19 -- long term effects
no. 8 -- social fairness	no. 22 -- comparative effects
no. 14 -- costs	no. 24 -- side effects
no. 17 -- adoptability	no. 26 -- quality of evidence
no. 18 -- strength of effect	

5. experts
6. information and products are reviewed
7. comparisons are made
8. uses scales
9. collection of individuals with calibration
10. evidence required

This is a very useful document discussing procedures of the review and some of the issues concerning the criteria. Includes information on how products scored on various criteria.

I. Sharing Educational Success

Reported by the National Advisory Council on Supplementary Centers and Services (16) and (17).

1. programs
2. mixed topic
3. to recommend
4. a. screening criteria:

   no. 1 -- priority of the need
   no. 2 -- do goals address need?
   no. 14 -- costs
   no. 17 -- adoptability
   no. 18 -- strength of effect

4. b. scaled criteria:

   no. 25 -- completeness of evidence
   no. 26 -- quality of evidence
   no. 27 -- extent to which evidence can be generalized

5. experts
6. programs themselves
7. no comparisons are made
8. uses scales
9. panel
10. evidence required

This review seems to us to be a prime example of a poorly designed review, which, because of its design, does not choose the programs it intends to choose. The purpose of the review is to choose Title III programs that were "outstanding" and "successful in terms of doing what they said they would do" (16, p. 5). But ratings are done solely on the quality of the evidence. This allows such results as giving high ratings to programs that have clear evidence of being unreasonably expensive; see the rating form in (17, p. 87). Effectiveness is rated only

indirectly, by a rating of the quality of the evidence that the program meets its objectives, allowing program directors possibly to list as objectives those things known to be attained, or to scale down the objectives to the level that is attained; see (17, p. 74, #4d). The significance of each objective is rated in terms of the evidence that the objective is "educationally significant," with the possible result that although insignificant objectives are screened, a program with good evidence of being mildly significant could receive a higher rating than one with fiar evidence of being very significant; see (17, p. 81, #6).

Much of the confusion results from using a verification procedure to make a choice among programs. Choices should generally be made by reviews, which consider the quality of the programs, the degree of effectiveness, and the significance of the objectives, as well as the quality of the evidence. A catalog based on a selection of verified programs can only be useful if it stresses that the programs included are <u>not</u> screened to be "outstanding," but just to have good evidence. And those catalogs should be sure to list what claims are verified, so that users can see immediately if the claims that are verified are of any significance.

J. Project Information Packages

Reported by RMC Research Corp. (21)

1. programs
2. special topic
3. to package
4. screening criteria:

no. 2 -- do goals address need?	no. 25 -- completeness of evidence
no. 14 -- costs	no. 26 -- quality of evidence
no. 17 -- adoptability	no. 27 -- extent to which evidence can be generalized

no. 18 -- strength of effect

no. 22 -- comparative effects

no. 28 -- comparative evidence

5. staff of review project
6. program itself is reviewed
7. comparisons are made
8. does not use scales
9. panel
10. evidence required

This is an especially well documented project designed to identify and package effective compensatory education programs in reading and mathematics. Several reports were submitted, covering the development of the criteria and the procedure for validating achievement gains, as well as other aspects of the project. The special focus of the project, and the fact that effectiveness in all cases can be measured by standardized tests, prevents their conclusions from being applicable to other kinds of reviews. It does contain excellent material for program reviews.

K. Technological Applications Project

Reported by the Instructional Development Division, U.S. International University (22).

1. programs, products, and practices
2. mixed topic
3. catalog review
4. screening criteria:

no. 17 -- adoptability

no. 18 -- strength of effect

no. 26 -- quality of evidence

5. staff of review project
6. only information is reviewed
7. no comparisons are made
8. does not use scales
9. panel
10. evidence sometimes used

A project to "inventory and classify" what are called "applications of instructional technology in educational settings" (22, p. 9). Little information is given on the review procedure used.

L. Panel Review of Products

Reported by the Educational Testing Service (23).

1. products
2. mixed topic
3. dissemination
4. scaled criteria:

no. 1 -- priority of need	no. 16 -- acceptability to users
no. 2 -- do goals address need?	no. 17 -- adoptability
no. 4 -- cumulative effects	no. 18 -- strength of effect
no. 5 -- product uniqueness	no. 19 -- long term effects
no. 6 -- social acceptability	no. 24 -- side effects
no. 13 -- development process	no. 25 -- completeness of evidence
no. 14 -- costs	no. 26 -- quality of evidence
no. 15 -- staff effort needed	

5. reviewed by project staff
6. product and information are reviewed
7. no comparisons are made
8. scales are used
9. panel
10. evidence required

This is one of the most useful documents available on issues of conducting a multi-product review. Full information is included about the criteria used, the rating form (developed by Scriven), and the procedures followed. The review is done entirely in terms of goals, with a strong emphasis on side effects due to this orientation. Our model rating form is largely based on Scriven's, with an additional criteria, especially in the area of intrinsic quality.

IX. Some Useful References

As with the previous section, this section is included with the hope that someone will follow up this report with further investigation of review procedures and criteria. The following are the articles and reports that we found to be extremely useful:

A. EPIE, EPIEform A (5). A very complete, thorough, and detailed rating form for obtaining descriptive and evaluative information. It includes sections on product description, product development, product usableness, goals and objectives, content organization, methodology, evaluation, and overall assessment and evaluation.

B. EPIE, "Evaluation Practices Used in the Selection of Educational Materials and Equipment." (6). A survey and brief analysis of procedures used in various

states for selecting educational materials and equipment. The emphasis is mainly on the procedures, and on who makes what decisions, with little attention paid to criteria used.

C. EPIE, "Improving Materials Selection Procedures: A Basic 'How To' Handbook." (7). A discussion of recommended procedures for selecting educational materials, designed for users. Includes a checklist of criteria to be considered.

D. EPIE, "Selecting among Textbooks." (8). A very useful report on procedures, criteria, and rating forms written for actual users, that is, people deciding on the actual materials to use in a class. It includes methods of analysis of the instructional setting and the instructional design of the textbooks, and a discussion of possible procedures to be followed. It describes two systems of descriptive analysis of texts and includes examples of forms for collecting information from users and for doing actual ratings.

E. Hood, "Evaluation of Instructional Materials for Educationl RDD&E Content Areas." (11). This paper is part of a project for the "Design of a R & D Information System." The paper discusses the issues involved in evaluating instructional materials both for acquisition in a collection and for dissemination. It also discusses differences between reviews done for different purposes: selection, classification, criticism, and evaluation. Two rating forms are recommended and analyzed, the ETS form (23) and the Eash Assessment Instrument.

F. Klein, "NIE Product Rating Form." (13). This rating form has the most thorough list of criteria we have seen, and is the source of our list of criteria and many of the ideas about them. Asks for a rating of each type of possible evidence for each criteria, but with no method of using these ratings in the final rating. Tends to be a little confusing.

G. Scriven, "Evaluation Perspectives and Procedures." (19). A very useful work discussing central issues of evaluation and review. Includes a "checklist

for the evaluation of products, producers, and proposals," and discussions about goal free evaluation, causal investigations, and cost analysis. Many of the ideas in our report come from this paper, or were provoked by it.

H. SWRL. "SWRL Instructional Product Selection Kit." This is an intriguing product, designed to be used in workshops to train people to select instructional products. The kit includes an instructor guide, booklets that discuss considerations in selecting products, and simulated program brochures that describe three elementary school science programs.

Assessment of products is based on a concept of a "complete instructional product," which seems to mean, basically, one with clearly stated, observable outcomes, with evidence to demonstrate that these are achieved and which contains a means of assessing the students' progress toward these outcomes. A few other considerations are also included, such as cost, adoptability, and product appeal.

While the kit may be an effective way to train people to pick "complete instructional products," it virtually ignores questions that are crucial in most user reviews, especially the need question of how to choose a product that does what is needed. A product that meeets all the criteria for being "complete" will not be the best product for a situation unless the objectives are very appropriate for that situation.

In general this kit is a good example of an attempt to make review activities more systematic and "scientific" by ignoring the important questions.

X. Conclusions and Recommendations

To repeat and summarize the main conclusions and recommendations that have been emphasized in this report:

(1) The need to include effectiveness criteria. Most reviews are based solely on choosing the product with the most appropriate contents. This is important too, of course, but we would like to see at least some attention to the demonstrated effectiveness of the products whenever useful evidence is available.

(2) The importance of not misusing evidence. This is a balance to the first point. When evidence, especially for effectiveness, is a major aspect of a review, it must be treated with a healthy scepticism. One must beware of evidence obtained so carelessly that it shows nothing. One must beware of evidence that does not show what needs to be shown. One must beware of evidence that proves that the goals of the product are achieved, but only because the goals are inappropriate or trivial. One must be sure to give a rating for whatever the evidence does show, as well as for the credibility of the evidence. And one must be careful to avoid the tendency to choose products or programs with the least significant goals, simply because they have the best evidence available.

It seems possible that once evidence is treated with this healthy scepticism, very little of it will be found to be useful. If this turns out to be the case, we must seriously consider eliminating or at least deemphasizing evidence as a factor in making choices between products, and once evidence is eliminated, the entire effectiveness area would have to be dropped also. This would suggest that some other means of increasing the quality of educational products be looked for, such as the learner verification and revision process.

(3) The review is only as good as the judgment of the reviewer. Reviewers should be given the opportunity to use their judgment as freely and as explicitly as possible. Their assumptions should be stated explicitly. They should make this the final summary ratings. Reviews cannot be done by computers, and they cannot be done by simple mathematical formulas. The job of the reviewer is too difficult for that, because of the necessity of reaching a decision on the basis of whatever information is available, no matter how incomplete.

(4) The importance of value judgments. A thorough review must involve a large number of value judgments, about the priorities of the needs, the relevance of the goals, the suitability of the instructional strategy, the attitude toward social fairness, and so on. To avoid these issues is to say in effect that they are not important; the result, then, may be low-priority products, with irrelevant goals and unsuitable instructional strategies, and with prejudiced attitudes towards social fairness. All these issues must be faced in a review. We believe that they should be faced head on, as explicitly, openly, clearly, rationally, and well thought out as possible.

(5) The need for future research on review procuedures. Several areas seem urgent to us. First is the trial use of the model rating form proposed in this paper and other detailed forms, in order to get feedback on their usefulness and on possible revisions, as well as comparative information. This would be a kind of L-BR of these forms. Second, a test of the reliability of various forms; unless one can get similar results from different reviewers one can have no confidence in the review procedure at all. Third, a large-scale study of the credibility of evidence available on various products in order to see how often the evidence is useful at all. And finally, a search for review models and for methods of evaluating reviews.

(6) The need for a more thorough coverage of criteria in reviews. Most reviews were found to be very incomplete in this respect. A concern for social fairness was noticeably lacking. Of all the reviews we considered, social fairness was only included in the NIE reviews (12) and (14).

Bibliography

(1) Automated Services, "Collecting, Editing, Preparation, and Publication of Information Concerning Promising Educational Programs and Practices." Final Report. Washington, D.C.: Automated Services, Inc., August, 1974.

(2) Bowers, J., P. Campeau, and A. Roberts, "Identifying, Validating, and Multi-Media Packaging of Effective Reading Programs." Final Report. Palo Alto: American Institutes for Research, Dec., 1974.

(3) Doughty, P., and D. Gooler, "A Summary of the Protocol Evaluation Conference." Syracuse: Development and Evaluation Associates, Dec., 1973.

(4) Dunn, J., L. Steel, J. Melnotte, and A. McCormick, "Effectiveness of Educational Laboratories and Centers: A System for the Evaluation of Educational Research and Development Products." Palo Alto: American Institute for Research, Feb., 1973.

(5) EPIE, EPIEform A. Instructional Materials Analysis Form. New York: EPIE Institute, 1974.

(6) EPIE, "Evaluation Practices Used in the Selection of Educational Materials and Equipment." Albany: The State Education Department, Division of Educational Communications, Dec., 1969.

(7) EPIE, EPIE Report: Number 54; Improving Materials Selection Procedures: A Basic "How To" Handbook. New York: EPIE Institute, 1973.

(8) EPIE, "Selecting among Textbooks." A Special Report for UNESCO. New York: EPIE Institute, Feb., 1974.

(9) Guba, E., and W. Gephart, "Training Materials for Research, Development and Diffusion Training Programs." Washington, D.C.: National Center for Educational Research and Development, USOE, 1970.

(10) Gunderson, D., "Protocol Materials." Washington, D.C.: National Center for the Improvement of Educational Systems, USOE, undated.

(11) Hood, P., "Evaluation of Instructional Materials for Educational RDD&E Content Areas." San Francisco: FWLERD, July, 1973.

(12) Klein, S., "An Autopsy of a Pilot Mini-Panel Review of Teacher Training Products in the Social-Emotional Domain." Washington, D.C.: NIE, Oct., 1974. (Draft)

(13) Klein, S., "NIE Product Rating Form." Draft 2. Washington, D.C.: NIE, Sept., 1974.

(14) Klein, S., M. Stump, and M. Dilworth, "Summary Report on the Product Quality Review Associated with the 1974 Dissemination Request for Proposals." Washington, D.C.: NIE, Teaching and Curriculum Program, August, 1974.

(15) McGraw-Hill, "Guidelines for Equal Treatment of the Sexes in McGraw-Hill Book Co. Publications." New York: McGraw-Hill Book Co. (no date given).

(16) National Advisory Council on Supplementary Centers and Services, "Innovative Educational Practices." Volume II. Washington, D.C.: NACSCS, Oct., 1974.

(17) National Advisory Council on Supplementary Centers and Services, "Sharing Educational Success -- A Handbook for Validation of Educational Practices." Washington, D.C.: NACSCS, undated.

(18) NEA, "Checklist for Selecting and Evaluating U.S. History Textbooks." Washington, D.C.: NEA, 1973.

(19) Scriven, M., "Evaluation Perspectives and Procedures." In Evaluation in Education: Current Applications, J. Popham, ed. Berkeley: McCutchan Publishing.

(20) SWRL, "SWRL Instructional Product Selection Kit." Los Alamitos, Calif.: SWRL, 1975.

(21) Tallmadge, G. K., "Development of Project Information Packages for Effective Approaches in Compensatory Education." Los Altos, Calif.: RMC Research Crop., Oct., 1974 (Report No. UR-254).

(22) Urbach, F., and T. Kepner, "Report for the Technological Applications Project." Corvallis, Ore.: Instructional Development Division, United States International University, June, 1974.

(23) Walton, W., B. Esser, M. Epstein, E. Margosches, and W. Schrader, "Selection of Exemplary Educational Products." Princeton: Educational Testing Service, April, 1973.

Addendum

(24) Epstein, M., E. Margosches, W. Schrader, and W. Walton, "Selection of Products for Focused Dissemination." Princeton: Educational Testing Service, Fall 1972.

Appendix A

Glossary

Credibility -- The overall rating of success of the evidence in supporting claims made for a product. All other factors, such as relevance and completeness, quality, extent to which evidence can be generalized, difficulty, and comparative evidence, would be considered in reaching this rating.

Criterion -- An item on which a product or program is to be judged. Scaled criterion: One on which a rating is to be given, such as on a 0 to 4 scale. Screening criterion: One on which a product is to be judged acceptable or nonacceptable.

Empirical Evaluation -- A study which gathers evidence on the success of a product in attaining its goals or objectives, or on the strength of effects of a product, with the help of tests given before and after the product is used, and with adequate controls.

Exemplary Product -- One judged to be of high quality in all aspects of the review, and to have adequate evidence supporting this judgment. This term is only useful if the criteria by which the product was judged are made explicit.

Goals -- The broad purposes of a product.

Instructional Product -- One designed with an educational purpose in mind.

Learner verification and revision -- The "empirical process of fine-grained data-gathering and analysis aimed at ascertaining, improving, and maintaining the teaching-learning effectiveness of a particular instructional material, beginning with its earliest developmental version and continuing through all revisions -- until such time as the material is no longer published" (5, unpaginated).

Needs -- The problem or problems addressed by a product.

Need Criterion -- Those criteria dealing with the importance or priority of the goals or effects of a product.

Objectives -- The specific effects a product is designed to have in order to accomplish its goals.

Practice -- Current usage: any procedure used in a specific location, including pedagogical techniques, lessons, courses or units within courses, programs, and administrative procedures. Proposed usage: an educational procedure, ongoing at a specific site, less comprehensive than a program.

Promising practice -- A practice tentatively believed to be exemplary, but without conclusive evidence yet available.

Product -- A tangible and transportable educational material, which is basically self-contained.

Program -- "An instructional system ongoing in a specific location" (2, p.1), usually including products that it uses, a staff, a budget, standard procedures, general goals and specific objectives.

Review -- An activity performed in order to make a choice among various alternative educational materials.

Validation -- Current usage: unclear. Proposed usage: the process of verifying that a product attains its objectives. This activity would only be useful if the objectives that were verified are made explicit when the product is described as validated, and if it is clear that the term "validated" means no more than that.

Verification -- The process of checking the accuracy and justifications for the statements made about a product. These statements may describe product costs, claims, assurances, etc.

Appendix B

Dimensions of Catalog Reviews

Catalog reviews comprise a subject that is deserving of separate study. We would recommend the following dimensions of catalog review activities be considered in any such study. These dimensions could also be used as items to be included in a catalog of catalogs.

A. Type of Screening. Catalogs use one or more types of screening processes:

(1) Subject area -- only products covering a particular subject matter are included in the catalog; no screening by quality is done. This is the most common type of catalog.

(2) Miscellaneous screening criteria -- some catalogs are only interested in products or programs that meet a few specified screens, such as being a Title III program, a programed or individualized textbook and so on.

(3) Verification -- only products successfully passing a verification stage are included. A catalog like this can only be useful if the verification process is made explicit, and if the catalog states for each product what it is that is verified. It is not enough to know simply that a product is verified; perhaps what is verified is that the product is unreasonably expensive, or that it attains goals that happen to be trivial.

(4) Quality -- only products judged to be of high overall quality are included. In effect, a complete review would have to be conducted to choose products for a catalog of this sort. Such a catalog would be most helpful for users, if they have confidence in the group conducting the review, and if full information about the review procedure and criteria are given.

B. Criteria. Catalog reviews can use various kinds of criteria in the screening process. These can be used as screening criteria which all must be passed by a product in order to be included in the catalog. Or they may be used as scaled criteria, in a process by which products are rated on various

criteria and then those with the highest ratings are included in the catalog. These criteria fall into seven categories, of which the first five are the same as for reviews in general:

(1) need

(2) effectiveness

(3) intrinsic quality

(4) practicality

(5) credibility of evidence

(6) subject matter

(7) specialized criteria -- Title III programs, programed textbooks, for example. Most catalogs use only the sixth kind, and maybe the seventh.

C. Information Included in the Catalog. The information included falls into four categories:

(1) Information on the selection process used. A catalog that has used any screening process at all should present full information about the procedure and criteria used; the conclusions of the process are not sufficient.

(2) Content information. Catalogs that contain full descriptive information about the contents of the products and about the reading level, emphasis, treatment, and viewpoint of the products are of great help to users who often have specific ideas about what they are looking for.

(3) Verification information. Catalogs can go a long way toward helping users find high quality products by verifying the claims and assurances made about a product and the evidence supporting these claims. Users then have some source of reliable information about the products. It is not enough to state in the catalog that the claims and assurances have been verified, or even that the evidence showing that the product meets its objectives has been verified. The claims, objectives, and assurances that have been verified must themselves

be included, especially because of the possibility that producers may tone down their objectives to match what can be demonstrated. Users need to see what it is the product can do, not just be told it can do what it is supposed to. Catalogs might also consider including information on the claims and assurances that cannot be verified in order to give a perspective on what is intended for the product but not realized.

(4) Quality information. Some catalogs give information about the judged quality of the product. This could be done by giving detailed results of a systematic review for each product, or by including letters by experts expressing their opinions of the products, similar to book reviews in a magazine. Quality information could be included in the catalog whether or not the screening process by which products are selected for inclusion involved quality criteria.

Appendix C

Operational Criteria for Social Fairness: Some References.

Some very useful beginnings have been made in developing operational criteria for social fairness. The most helpful reports that we have found are:

(1) Curriculum Framework and Instructional Materials Selection Unit, "Guidelines for Evaluation of Instructional Materials for Compliance with Content Requirements of the Educational Code." Sacramento: Calif. State Board of Education, Sept. 12, 1974.

-- Includes criteria with respect to female and male roles, and ones with respect to ethnic and cultural groups.

(2) Ginn and Company, "Editorial and Graphics Criteria." Lexington: Mass. Ginn and Co., Publishers, Jan. 29, 1974.

-- Includes criteria for depicting the sexes and for depicting members of minority groups in textbooks.

(3) NEA, "Checklist for Selecting and Evaluating U.S. History Textbooks." Washington, D.C.: NEA, 1973.

-- Guidelines to improving treatment of minorities and women in textbooks.

(4) McGraw-Hill, "Guidelines for Equal Treatment of the Sexes in McGraw-Hill Book Co. Publications." New York: McGraw-Hill Book Co. (no date given).

Appendix D

Legislation Concerning the Selection of Educational Material: Some References.

We have found the following reports to be useful in this area.

(1) EPIE "Pilot Guidelines for Improving Instructional Materials through the Process of Learner Verification and Revision." New York: EPIE Institute, June, 1975.

(2) Geffert, H., R. Harper, and D. Schember, "Textbooks and Other Educational Materials: Improving Effectiveness, Dealing with Bias, Selecting for Use in the Public Schools: A Model Statute." Washington, D.C.: Lawyers' Committee for Civil Rights under Law, Jan., 1975.

(3) Pearsin, J., "A Handbook of State Laws and Policies Affecting Equal Rights for Women in Education." Denver: Education Commission of the States, March, 1975. Report #62 from the Equal Rights for Women in Education Project, Ford Foundation Grant.